Journal of Official Statistics
vol. 36, 2 (Jun 2020)

sciendo

Journal of Official Statistics, Vol. 36, No. 2, 2020, pp. 229–235, http://dx.doi.org/10.2478/JOS-2020-0012

Letter to the Editors

COVID-19: Unprecedented Situation, Unprecedented Official Statistics

The COVID-19 outbreak dominated the beginning of 2020. Almost every country and every socio-economic sector is facing this unique situation. Official statistics need to confront new challenges, both internally and externally. Internally, every National Statistics Office (NSO) needs to protect their workers, reorganize their way of working, and ensure the regular statistics production. Externally, NSOs are called to make a statistical description of an unprecedented complex reality.

The aim of an NSO is not only to produce official statistics but also to provide a realistic 'picture' of our world, even in times of crisis. However, an official statistics system should not limit itself to those standard tasks. NSOs should provide and participate to the production of data and information on COVID-19. For instance, the United Nations Statistics Division is actively recommending guidelines, sharing experiences and collecting data for COVID-19 (United Nations Statistics Division 2020). However, under the threat of COVID-19 many countries are considering official statistics as a non-essential service (Cheung 2020a). Are official statistics a non-essential service?

The aim of this letter is: (1) to analyze the role of NSOs during the COVID-19 health crisis, (2) to summarize challenges and opportunities lying ahead, (3) to trigger the discussion about the role of NSOs.

For the sake of simplicity, let us distinguish between two different temporal phases of the COVID-19 epidemic. The first phase is predominated by health crisis and lockdown. This phase affects the population. The second phase starts when the health crisis is under control (IAOS 2020). Once the lockdown has been revoked, the country is facing a new socio-economic situation that was completely unpredictable in January 2020. This last phase is full of challenges and opportunities for NSOs.

During lockdown, citizens are reading or listening to the figures of new COVID-19 cases and daily deaths. These numbers are unreliable, not harmonized and scarce. They do not provide a secure starting point to understand the situation, nor do they help in making proper decisions. What drives the production of these figures – cases and deaths of COVID-19 – is a medical objective. These figures are necessary to diagnose and care for patients. Nevertheless, we also need figures to better understand the spread of COVID-19. We need figures to help implement the lockdown only when and where lockdowns are necessary. After the health crisis, we will need statistics to help us fully understand the new socio-economic situation and the details of the economic recession. We need reliable figures to plan for the future.

Acknowledgments: The author would like to thank Moshe Kim, Dario Cattivo, Barbara Di Gennaro Splendore, Leonard Schembri and Sonia Judith for their comments and suggestions. The views expressed in this letter are exclusively those of the author.

1. Official Statistics During the Health Crisis

Under the threat of COVID-19, the priority of NSOs is the health of its workers. The second priority is to describe the country's situation based on statistical methodology and data collection. Policymakers and citizens need official statistics to make informed choices to manage the health crisis.

In the media, two figures are predominant: number of deaths and number of cases of COVID-19. From the point of view of an NSO, these data are produced by the health system and they are administrative data. We can call these figures "medical figures". We can consider the production of these data as a side effect of the excellent work conducted by doctors and nurses who care for COVID-19 patients. The aim of medical figures is not to describe the spread of COVID-19 or the overall socio-economic implications of the outbreak. Unfortunately, figures produced by health authorities cannot provide crucial information. We need to know how, when and why COVID-19 spreads among the population. For instance, we need to know how many asymptomatic carriers of COVID-19 there have been in our country. Only a scientific random sample could give us this information (see Ioannidis 2020; Di Gennaro Splendore 2020; Alleva et al. 2020).

Moreover, since medical figures – the number of deaths and cases of COVID-19 – are not produced by or on behalf of the NSO, the NSO cannot assess the quality of these data. While not referring to COVID-19, Radermacher (2020) has proposed that the statistical system could take on the assessment, management and certification of data. For instance, the German Federal Statistical Office declared that it does not collect real-time data on the outbreak. When looking for data on COVID-19, citizens are redirected to a different webpage. Interestingly, COVID-19 is not present on the webpages of several NSOs (Figure 2 in Misra et al. 2020).

An important issue about the number of deaths of COVID-19 is the interpretation of the main cause of death. From this perspective, figures of COVID-19 deaths are problematic. To make an analogy, if you have the flu and a train runs you over, you would have died because of the train not because of the flu. The number of COVID-19 deaths should only indicate the number of deaths for which COVID-19 is the main cause of death. Instead, different countries apply different definitions and different criteria. Often, various regions within the same country apply different criteria. Normally, the national statistical systems work on metadata and harmonized definitions. Italy, Ecuador and the United States (CDC 2020) report COVID-19 deaths where the main cause of death is not COVID-19. The media compares deaths in 2020 with previous years' deaths. These are unreliable data, as demographers well know. To have a comprehensive analysis of these data we should include an analysis of age groups.

If you do not take a blood test, then you are not identified as having, for example, high cholesterol. The number of cases of COVID-19 depends on how many people were tested in each country and also how they were tested. Availability of testing and different types of tests should be reflected in the statistics. Likewise, data should include people in hospitals, asymptomatic cases and people who came in contact with infected people. The number of COVID-19 cases we have today is not a reliable measure of how many people are infected by the disease. These numbers are useless both for comparison among countries and over time. The figures of COVID-19 cases we have are an underestimation of true figures.

Because of this underestimation of the number of COVID-19 cases, in the media, we find systematic overestimations of the Case Fatality Rate (CFR). Since the majority of countries do not test the whole population, figures are greatly biased. If you cannot measure the main variable of the health crisis, you cannot manage the crisis and its implications.

Due to their data collection capacity, infrastructure and experience, NSOs could support and help health authorities by producing reliable data. Statistical systems should provide data collection support, quality control of figures and appropriate communication of statistics. The NSOs could be matching COVID-19 cases with socioeconomic aspects (like gender, age, income, etc.), previous medical problems, address (GPS), and so on. Having those multivariate data could give the possibility to use more sophisticate statistical models. Statistical systems are key to implementing real-time standardized reporting of the results and disaggregated data, and thus help assess the implications of COVID-19.

How can we acquire information on COVID-19? To the best of our knowledge, there are only two possibilities to obtain this information: either via a census of the population or via a random sample representative of the population. In most countries, a census is not practicable. We need a random sample representation of the population. Different sample designs and different possibilities can be implemented. Every person in the sample who is tested for COVID-19 also needs to answer a questionnaire. The questionnaire would include questions about the clinical evaluation, socio-demographic characteristics, personal characteristics, housing characteristics, and lifestyle of the individual. Additional relevant information can be obtained – in many countries – through administrative registers.

Who can access more information on COVID-19? NSOs are expert institutions on population, sampling and data collection. They are in charge of censuses, employment surveys, as well as many other surveys. The collaboration between NSOs and National Health Systems at the national level could guarantee the necessary expertise to implement a random sample of the population under the threat of COVID-19. As a result of the preparation for the Population and Housing Census, every NSO is ready to investigate its own population. Between the years 2020 and 2021, almost all NSOs in the world will implement the Census of Population and Housing. It is likely that almost all NSOs already have the master frame ready to prepare the census. This could serve as a starting point to implement a random sample. Between 2005 and 2014, more than six billion people around the world – more than 90% of the world's population – were enumerated by population censuses. Only 21 countries did not conduct a census (United Nations Fund for Population Activities 2016).

The United Nations Fundamental Principles of Official Statistics, states that: 'Official statistics provide an indispensable element in the information system of a democratic society, serving the government, the economy and the public with data about the economic, demographic, social and environmental situation' (United Nations 1994). Heinrich Bruengger, former Director of the Statistical Division of UNECE, explains the same principle as follows: 'The purpose of official statistics is to produce and disseminate authoritative results designed to reliably reflect economically and socially relevant phenomena of a complex and dynamic reality in a given country' (Bruengger 2008). The mission of the NSOs is to inform us, the people. Lockdowns in many countries are very restrictive: presently we are experiencing restriction of freedom of movement and restriction of economic freedom. The necessity of such strict measures is why decision-makers and citizens need reliable information about COVID-19. 'In a democratic society

the independence of official statistics has the same status as the freedom of speech for the citizens.' (Jeskanen-Sundström 2007, 1).

To recapitulate, the NSOs during the crisis could:

- help with data collection regarding the crisis,
- explain, manage and/or certify data and statistics, and
- support a random sample representative of the population to identify the spread of the outbreak.

2. Official Statistics After the Health Crisis

In this section, we revisit the unprecedented challenges and remarkable opportunities posed by the health crisis to NSOs.

If we consider the tourism sector or airline companies, we immediately realize that the figures in the next official statistics will be much lower than what we expected in December 2019. By now, our forecasts for 2020 have lost all their meaning. For example, can the price of a flight ticket serve to calculate the inflation? Many services, such as package holidays and sports event tickets are no longer on the market. As of today, we do not know the price of several products. Inaccurate statistics lead to wrong decisions concerning measures to support the economy. Despite all the difficulties, NSOs need to provide accurate and reliable statistics. Therefore, it is necessary to bridge statistical gaps between present data and post-COVID-19 data to understand all the economic implications of the pandemic.

(1) Below are a few examples of the impact that the COVID-19 crisis has had on the bodies of official statistics. Almost all areas of statistical production have problems, each presenting different levels of criticality.
   - Response rates to surveys might decrease. For example, in some cases, survey interviews were canceled for several weeks (Istat Italy 2020),
   - Official statistics originating from past economic interrelations are no longer valid. The estimation procedures cannot provide reliable results in this special circumstance. Therefore, real data are all the more important for evaluating the present situation. The values under COVID-19 in the time series are informative outliers and not atypical values (Eurostat 2020),
   - Governments decided on massive amounts of social benefits and unprecedented subsidies for production. The classification and reporting of these new economic policies need to be understood and coordinated (Eurostat 2020), and
   - The postponement or cancellation of some releases and publications is also problematic. For instance, the 2020 Population and Housing Census will be shifted (Cheung 2020b). Some NSOs advise possible rescheduling of the release calendar (INE Spain 2020; NSO Malta 2020).

(2) However, this crisis offers a number of opportunities for NSOs:
   - As of today, up-to-date and short analyses are more appreciative and useful than large studies. The latter provide relevant but overdue insights. In order to provide information that supports decision-makers and citizens, official

statistics need to be as reactive as possible with regard to the present situation and real evolution of social and economic dynamics caused by the COVID-19 pandemic. Is the quarterly GDP enough? A higher frequency release for the GDP and other socio-economic indicators, in this significant crisis, could be an extra tool to help decision-makers and citizens,

- Adequate granular sources different from traditional ones can help pinpoint emerging concentrations of needs. Additionally, they can measure extra-ordinary changes in real-time. Already before the COVID-19, we could read: 'Official statistics are fundamental to democracy. With increasing demands for more relevant, frequent and rich statistical information, and declining resources, national statistics offices are continually looking for more cost-effective ways in the production of official statistics' (Tam and Kim 2018),
- Social distancing gave the definitive impulse to pass from CAPI (Computer-assisted Personal Interviewing) to CAWI (Computer-Assisted Web Interviewing). The NSOS had already questioned CAPI or PAPI (Paper Assisted Personal Interview), and were slowly moving to CATI (Computer-Assisted Telephone Interviewing). Nevertheless, this could be the moment to implement CAWI. This is only one aspect of digitalization that NSOs need to face moving forward. In addition, it would be the right moment to use machine learning and web-scraped data (e.g., in the case of on-line prices),
- In a time of crisis, fake news is especially insidious and often faster than real news. Statistical Offices should manage and certify the quality of statistics produced from outside sources in order to include them in official statistics (Radermacher 2020),
- NSOs are in the best position to meet policymakers' rising demand for information about health services. This information will be crucial to effectively manage the consequences of the epidemic, forecast and set up a system of prevention and quick response, and
- Finally, this is the appropriate moment to shift the attention of official statistics from offer to demand.

These challenges can be an opportunity to reinvent and reinforce the role of NSOs all over the world. New tools and strategies should be ready for the next pandemic, the climate crisis or economic recession.

3. Moving Forward

The prolonged lifespan of the COVID-19 pandemic provides NSOs with challenges that could not have been predicted in January 2020. This unprecedented situation is described by unprecedented official statistics and probably by the unprecedented quality of official statistics. Statistical Offices have to collaborate sharing experiences and ideas among themselves (United Nations Statistics Division 2020). So far, few countries and few NSOs have had the capacity to singlehandedly tackle these new issues, challenges and opportunities.

During the health crisis – or any another crisis – NSOs have the chance to share their expertise in sampling design, data collection and data quality. These assets are crucial to fully understand all the details of the crisis and its implications. After the crisis, the world

will face a large and deep socio-economic recession with an unforeseen and unpredicted lack of data and information. Official statistics have a key role for decision-makers and citizens. The NSOs have exceptional challenges and extraordinary opportunities to redesign their roles and their tools.

The COVID-19 crisis can convert official procedures, and inflexible routines in up-to-date analysis, modeling, experimental statistics, digitalization, and complementary adequate granular sources. More than ever, statistical infrastructure and methodological expertise represent a vital resource. Information is critical for political and economic decisions. Statistical systems must provide reliable statistics to manage the crisis, but they also need to learn from the crisis.

4. References

Alleva, G., G. Arbia, P.D. Falorsi, and A. Zuliani. 2020. "A sample approach to the estimation of the critical parameters of the SARS-CoV-2 epidemics: an operational design with a focus on the Italian health system." Research Gate. Available at: https://www.researchgate.net/publication/340514422_A_sample_approach_to_the_estimation_of_the_critical_parameters_of_the_SARS-CoV-2_epidemics_an_operational_design_with_a_focus_on_the_Italian_health_system (accessed April 2020).

Bruengger, H. 2008. "How Should a Modern National System of Official Statistics Look?." UNECE, Statistical Division.

CDC, Centers for Disease Control and Prevention. 2020. Vital Statistics Reporting Guidance.

Cheung, P. 2020a. "Impact of COVID-19 on Official Statistics (2) - Is Official Statistics Non-Essential Service." Available at: https://www.linkedin.com/pulse/impact-covid-19-official-statistics-2-non-essential-paul-cheung/?trackingId=AWBHyvCKQqWOsV4f98T22Q%3D%3D (accessed April 2020).

Cheung, P. 2020b. "Impact of COVID-19 on Official Statistics." Available at: https://www.linkedin.com/pulse/impact-covid-19-official-statistics-paul-cheung/ (accessed April 2020).

Di Gennaro Splendore, L. 2020. "Random testing, quality of data and lack of information: COVID-19." Available at: https://medium.com/data-policy/random-testing-quality-of-data-and-lack-of-information-covid-19-a6e09a398d1d (accessed April 2020).

Eurostat. 2020. Website. *COVID-19: support for statisticians*. Available at: https://ec.europa.eu/eurostat/data/metadata/covid-19-support-for-statisticians (accessed April 2020).

IAOS, International Association for Official Statistics. 2020. *Official Statistics in the context of the COVID-19 crisis*. Website. Available at: https://officialstatistics.com/news-blog/crises-politics-and-statistics (accessed April 2020).

INE, Spain. 2020. Website. Comunicado relativo a la actividad del INE ante la emergencia sanitaria con motivo del COVID-19. Available at: https://www.ine.es/ine/comunicado1_ine_covid19.pdf (accessed May 2020).

Ioannidis, J. 2020, "A fiasco in the making? As the coronavirus pandemic takes hold, we are making decisions without reliable data." Available at: https://www.statnews.com/2020/03/17/a-fiasco-in-the-making-as-the-coronavirus-pandemic-takes-hold-we-are-making-decisions-without-reliable-data/ (accessed April 2020).

Istat, Italy. 2020. Website. *Informazioni dall'Istat nell'emergenza sanitaria*. Available at: https://www.istat.it/it/archivio/239854 (accessed May 2020).

Jeskanen-Sundström, H. (2007). "Independence of Official Statistics, a Finnish Experience." Seminar on Evolution of National Statistical Systems. United Nations Statistical Commission, New York, 23 February 2020. Available at: https://unstats.un.org/unsd/dnss/docViewer.aspx?docID=1590 (accessed April 2020).

Misra, A., J. Schmidt, and L. Harrison. 2020. *Combating COVID-19 with Data: What Role for National Statistical Systems?* PARIS 21 - New Policy Brief.

NSO, Malta. 2020. Website. COVID-19 and the production of statistics. Available at: https://nso.gov.mt/en/nso/Pages/news/COVID19-information-notice.aspx (accessed May 2020).

Radermacher, W.J. 2020. *Official Statistics 4.0*. Springer.

Tam, S. and J. Kim. 2018. "Big Data ethics and selection-bias: An official statistician's perspective." *Statistical Journal of the IAOS* 34(4): 577–588. DOI: https://doi.org/10.3233/SJI-170395.

United Nations. 1994. *Nations Fundamental Principles of Official Statistics*. Available at: https://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx (Accessed April 2020).

United Nations Fund for Population Activities. (2016). *Annual Report*. Available at: https://www.unfpa.org/annual-report-2016 (Accessed April 2020).

United Nations Statistics Division. 2020. Website. *COVID-19 Response – Resources for Official Statisticians*. Available at: https://covid-19-response.unstatshub.org/ (Accessed April 2020).

Luca Di Gennaro Splendore

Institute of Tourism Studies Malta
Aviation Park, Aviation Avenue,
Hal Luqa Malta LQA 9023, Malta
Email: luca.di.gennaro.splendore@gmail.com

sciendo

Journal of Official Statistics, Vol. 36, No. 2, 2020, pp. 237–249, http://dx.doi.org/10.2478/JOS-2020-0013

Confidence Intervals of Gini Coefficient Under Unequal Probability Sampling

Yves G. Berger[1] *and İklim Gedik Balay*[2]

We propose an estimator for the Gini coefficient, based on a ratio of means. We show how bootstrap and empirical likelihood can be combined to construct confidence intervals. Our simulation study shows the estimator proposed is usually less biased than customary estimators. The observed coverages of the empirical likelihood confidence interval proposed are also closer to the nominal value.

Key words: Bootstrap; empirical likelihood; inclusion probability; survey weight; sampling design.

1. Introduction

Gini's (1914) coefficient is a widely used indicator for measuring income inequality in a wide range of area of economics and finance (e.g., Koshevoy and Mosler 1997; Ogwang 2000; Gajdos and Weymark 2005). The Gini coefficient is defined as the ratio of the area that lies between the 45° line and the Lorenz's (1905) curve given by

$$\mathcal{L}(x) := \frac{1}{E(Y)} \int_0^x y \, dF_Y(y), \tag{1}$$

where $F_Y(\cdot)$ is the cumulative distribution function of a positive random variable Y, and $E(Y)$ is the expectation of Y. An excellent review of various formulations of the Gini coefficient can be found in Giorgi and Gigliarano (2017).

Surveys are usually used to estimate the Gini coefficient. However, sampled units are rarely selected independently with equal probability, because of sample selection, which involves stratification and unequal probabilities. Two customary estimators for unequal probability sampling can be found in the literature (e.g., Langel and Tillé 2013, for a review). They are defined by Equations (13) and (14) in Section 5. The proposed estimator is different and based on a ratio, which allows to express it as an empirical likelihood estimator. Single stage designs are considered in this article. The proposed approach can be extended for multi-stage designs by using Berger's (2018a) approach.

Variance estimation of the Gini coefficient has been widely studied in the literature (e.g., Nair 1936; Hoeffding 1948; Glasser 1962; Sendler 1979; Beach and Davidson 1983;

[1] University of Southampton, Southampton, SO17 1BJ, UK. Email: Y.G.Berger@soton.ac.uk
[2] Ankara Yıldırım Beyazıt University, Esenboga Külliyesi Faculty of Business, Esenboğa, Ankara, 06760, Turkey. Email: iklimgdk@gmail.com

Gastwirth and Gail 1985; Schezhtman and Yitzhaki 1987; Sandström et al. 1985, 1988; Nygård and Sandström 1989; Yitzhaki 1991; Shao 1994; Binder and Kovaćević 1995; Bishop et al. 1997; Karagiannis and Kovaćević 2000; Ogwang 2000; Giles 2004; Modarres and Gastwirth 2006; Davidson 2009). Yitzhaki (1991) and Qin et al. (2010) proposed a variance estimator under stratified random samples. Asymptotic variance under stratified and clustered survey data can be found in Bhattacharya (2007). Berger (2008) proposed a jackknife variance estimator under unequal probability sampling. Langel and Tillé (2013) provided a comprehensive literature review on variance estimation for the Gini coefficient.

Sandström et al. (1988) have developed a confidence interval for the Gini coefficient based on normal approximation. Mills and Zandvakili (1997) consider bootstrap methods to compute interval estimates for the Gini coefficient. Qin et al. (2010) proposed pseudoempirical likelihood confidence intervals for the Gini coefficient under simple random sampling, using bootstrap and empirical likelihood methods. Qin et al.'s (2010) approach requires estimating the distribution function, and is not designed for unequal probability sampling. Other empirical likelihood intervals with independent and identically distributed observations can be found in Peng (2011). Empirical likelihood confidence intervals are range preserving; that is, the lower bound and the upper bound cannot be outside the parameter space [0, 1] of the Gini coefficient. The bounds are driven by the distribution observed from the data, rather than an asymptotic distribution. Empirical likelihood also offers the possibility of using some auxiliary information, which may improve the estimation of the Gini coefficient (Berger and Torres 2016). A review of empirical likelihood under unequal probability sampling can be found in Berger (2018b). Note that the confidence intervals proposed do not require an effective sample size or a design effect, unlike the pseudoempirical likelihood approach (Wu and Rao 2006) for unequal probability sampling.

In Section 2, we define the Gini coefficient. The proposed estimator is defined in Section 3. In Section 4, we show how bootstrap and empirical likelihood can be combined to construct confidence intervals. The empirical likelihood confidence intervals have the advantage of having bounds within the range of the Gini coefficient. Linearisation will not be required for empirical likelihood confidence intervals. Our simulation study in Section 5 shows that the proposed estimator can be more efficient than the customary estimator (e.g., Berger 2008; Langel and Tillé 2013). The coverages of the proposed empirical likelihood confidence interval are usually not significantly different from the nominal value.

2. The Gini Coefficient

Let $Y \geq 0$ denote a positive random variable with a distribution function $F_Y(y)$. The Gini coefficient is defined by

$$G_0 := \frac{2}{E(Y)} \int_0^\infty y F_Y(y) dF_Y(y) - 1 = 1 - \frac{1}{E(Y)} \int_0^\infty \{1 - F_Y(y)\}^2 dy. \tag{2}$$

Yitzhaki (1998) proposed an alternative expression of G_0 based on the minimum

$$Z := \min\{Y_a, Y_b\}$$

of two independent copies Y_a and Y_b of Y. Since $Z \geq 0$, we always have that $E(Z) = \int_0^\infty \{1 - F_Z(z)\} dz$, where $F_Z(z)$ denotes the cumulative distribution of Z.

Furthermore, since Z is the minimum of two random variables with the same distribution, we have that $F_Z(z) = 1 - \{1 - F_Y(z)\}^2$. This implies $E(Z) = \int_0^\infty \{1 - F_Y(z)\}^2 dz$. Thus, Equation (2) gives Yitzhaki's (1998) alternative expression (see also Peng 2011),

$$G_0 = 1 - \frac{E(Z)}{E(Y)}. \tag{3}$$

Let U be a finite population of N units, where N is a fixed quantity that is not necessarily known. Consider that we have N independent copies $\{Y_i : i \in U\}$ of Y. Let $\{y_i : i \in U\}$ be the realisation of these copies.

The empirical equivalent of $E(Z)$ is therefore

$$\bar{y}_U^* := \frac{1}{N(N-1)} \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} \min\{y_i, y_j\} = \frac{1}{N} \sum_{i \in U} y_i^*,$$

where

$$y_i^* := \frac{1}{N-1} \sum_{\substack{j \in U \\ j \neq i}} \min\{y_i, y_j\}.$$

Thus, the empirical version of Equation (3) is the finite population parameter

$$G_U := 1 - \frac{\bar{y}_U^*}{\bar{y}_U}, \tag{4}$$

where

$$\bar{y}_U := \frac{1}{N} \sum_{i \in U} y_i. \tag{5}$$

3. Estimation of the Gini Coefficient

Suppose that a sample S is randomly selected from U. We observe the values y_i for the sampled units $i \in S$. We shall use Neyman's (1938) design-based approach; that is, the sampling distribution is conditional on $\{y_i : i \in U\}$ and driven by the random selection of S. Thus, the values $\{y_i : i \in U\}$ and the parameter G_U will be treated as constants.

We consider that the population U is broken up into disjoint strata $U_1, \ldots, U_h, \ldots, U_H$ and $\cup_{h=1}^H U_h = U$. Within each stratum U_h, a sample of n_h units is selected with-replacement with unequal selection probabilities P_i, where $\sum_{i \in U_h} P_i = 1$. Let $\pi_i = n_h P_i$, when $i \in U_h$. Let S_h be the set of n_h labels for stratum U_h, where $S = \cup_{h=1}^H S_h$ and $n = \sum_{h=1}^H n_h$. We assume that we have a with-replacement or without-replacement sampling design with negligible sampling fractions, in order to justify the bootstrap approach. Fortunately, in practice, the Gini coefficient is estimated from social surveys, which are often based on negligible sampling fractions. The negligible sampling fraction is only needed for variance estimation and confidence intervals.

The estimator proposed for Equation (4) is

$$\hat{G}_\pi := 1 - \frac{\bar{y}_\pi^*}{\bar{y}_\pi}, \tag{6}$$

where $\bar{y}_{\pi}^{*}$ and $\bar{y}_{\pi}$ denote Hájek's (1971) estimators given by

$$\bar{y}_{\pi} := \frac{1}{\hat{N}} \sum_{i \in S} \frac{y_i}{\pi_i},$$

$$\bar{y}_{\pi}^{*} := \frac{1}{\hat{N}} \sum_{i \in S} \frac{\hat{y}_i^{*}}{\pi_i},$$

$$\hat{y}_i^{*} := \frac{1}{\hat{N} - \pi_i^{-1}} \sum_{\substack{j \in S \\ j \neq i}} \frac{1}{\pi_j} \min\{y_i, y_j\},$$

$$\hat{N} := \sum_{i \in S} \pi_i^{-1}.$$

The advantage of Equation (6) is the fact that it does not involve the estimation $F_Y(y)$. Note that Equation (6) reduces to Yitzhaki's (1998) under simple random sampling with a single stratum (see also Peng 2011; Giorgi and Gigliarano 2017).

Rescaled bootstrap (Rao et al. 1992; Rust and Rao 1996) can be used for variance estimation. This method is based on bootstrap weights (Rust and Rao 1996), given by

$$w_i^{(b)} := \frac{r_i n}{\pi_i (n-1)} \tag{7}$$

where r_i is the number of times i-th unit is selected by bootstrap. The variance between the bootstrap replicates can be used as a variance estimate. A bootstrap confidence interval based on the bootstrap quantiles can be derived (the "*histogram approach*").

The theory of bootstrap is well established, and little needs to be added. However, empirical likelihood is a new emerging topic, and little has been done on empirical likelihood confidence intervals for Gini, under unequal probability sampling. Peng's (2011) approach assumed an independent and identically distributed setting. Qin et al.'s (2010) method is based on simple random sampling. In Section 4, we show how an empirical likelihood confidence interval can be constructed with unequal probability sampling, in conjunction with bootstrap.

4. Empirical Likelihood Confidence Intervals

In this section, we show how Berger and Torres's (2016) approach can be combined with bootstrap. Empirical likelihood is based on estimating equations. It can be shown that Equation (6) is the solution to

$$\sum_{i \in S} \frac{1}{\pi_i} e(y_i, \hat{y}_i^{*}, G) = 0. \tag{8}$$

where

$$e(y_i, \hat{y}_i^{*}, G) := y_i(G-1) + \hat{y}_i^{*}. \tag{9}$$

By substituting Equation (9) within Equation (8), we obtain $\sum_{i \in S} \pi_i^{-1} y_i (G-1) + \sum_{i \in S} \pi_i^{-1} \hat{y}_i^{*} = (G-1)\hat{N}\bar{y}_{\pi} + \hat{N}\bar{y}_{\pi}^{*} = 0$. The solution to the last equation is indeed Equation (6).

Berger and Torres's (2012, 2014, 2016) "*empirical log-likelihood function*" is defined by

$$\ell_{\max}(G) := \max_{p_i : i \in S} \left\{ \sum_{i \in S} \log(p_i) : p_i > 0,\ \sum_{i \in S} \frac{p_i}{\pi_i} e(y_i, \hat{y}_i^*, G) = 0,\ \sum_{i \in S} p_i \boldsymbol{\delta}_i = \frac{\vec{\boldsymbol{n}}}{n} \right\}, \quad (10)$$

where G denotes a value within the parameter space, $\boldsymbol{\delta}_i$ is the vector of stratification variables defined by

$$\boldsymbol{\delta}_i := (\delta_{i1}, \ldots, \delta_{ih}, \ldots, \delta_{iH})^\top$$

and $\vec{\boldsymbol{n}}$ is the strata allocation given by

$$\vec{\boldsymbol{n}} := \sum_{i \in S} \tilde{\boldsymbol{\delta}}_i = (n_1, \ldots, n_h, \ldots, n_H)^\top.$$

Within Equation (10), we have two types of constraints. The constraint involving G is a moment condition which contains the standard sampling weights π_i^{-1}. We also have a stratification constraint $\sum_{i \in S} p_i \boldsymbol{\delta}_i = \vec{\boldsymbol{n}} n^{-1}$, which is not motivated by moment conditions. Equation (10) reduces to Owen's (1988) empirical log-likelihood function when we have a single stratum and $\pi_i = n/N, \forall i \in U$. The advantage of Equation (10) is that it can be used as a standard likelihood function for design-based inference. Note that Equation (10) differs from Peng's (2011) approach, even with a single stratum and $\pi_i = n/N$, because Peng's (2011) approach is based on splitting the sample randomly into two sub-samples of same size.

The "*maximum empirical likelihood estimator*" $\hat{G}_{EL}$ is defined as the quantity that maximises $\ell_{\max}(G)$. Berger and Torres (2016) show that this implies that $\hat{G}_{EL}$ is the solution to Equation (8). Thus, $\hat{G}_{EL} = \hat{G}_\pi$.

The empirical likelihood approach can also be used for confidence intervals based upon Equation (6). Consider the "*empirical log-likelihood ratio statistic*"

$$\hat{r}(G) := 2\left\{\ell_{\max}(\hat{G}) - \ell_{\max}(G)\right\}. \quad (11)$$

Berger and Torres (2016) showed that the empirical log-likelihood ratio statistic converges to an ancillary quadratic form, when $G = G_0$. Unfortunately, this quadratic form will not necessarily converge to a χ^2-distribution, because the $\hat{y}_i^*$ are estimated. In other words, this quadratic form is an ancillary statistic with an unknown distribution. We shall approximate this distribution using bootstrap.

In order to compute a α-level confidence interval, we would need to know the upper α-quantile of the distribution of $\hat{r}(G_0)$. This quantile can be approximated by the bootstrap distribution. Consider the rescaled bootstrap sampling weights given by Equation (7). Let $\hat{r}(G)^b$ be the b-th bootstrap value of Equation (11) based on bootstrap sampling weights given by Equation (7), with $G = \hat{G}_\pi$. The α-level bootstrap confidence interval is

$$\left[\min\{G : \hat{r}(G) \leq r_\alpha\};\ \max\{G : \hat{r}(G) \leq r_\alpha\}\right], \quad (12)$$

where r_α is the α-quantile of $\{\hat{r}(\hat{G}_\pi)^1, \ldots, \hat{r}(\hat{G}_\pi)^b, \ldots, \hat{r}(\hat{G}_\pi)^B\}$. Note that $\hat{r}(G)$ is a convex non-symmetric function with a minimum at $G = \hat{G}_\pi$. This interval can be found by using any root search method, such as Brent (1973) and Dekker's (1969) method, since the

bounds are the two roots of $\hat{r}(G) - r_\alpha = 0$. This can be achieved numerically by calculating $\hat{r}(G)$ for several values of G.

The empirical likelihood confidence intervals cannot be disjoint because $\hat{r}(G)$ is always convex, because of the strict concavity of the function $\sum_{i \in S} \log(p_i)$ within Equation (10).

5. Simulation Studies

Two customary estimators can be found in the literature (e.g., Berger 2008; Langel and Tillé 2013). They are given by

$$\hat{G}_\pi^{(1)} := \frac{2}{\hat{N}\bar{y}_\pi} \sum_{i \in S} \frac{y_i}{\pi_i} \hat{F}_\pi(y_i) - 1, \tag{13}$$

$$\hat{G}_\pi^{(2)} := \frac{1}{2\hat{N}^2\bar{y}_\pi} \sum_{i \in S} \sum_{j \in S} \frac{1}{\pi_i \pi_j} |y_i - y_j|, \tag{14}$$

where

$$\hat{F}_\pi(y_i) := \frac{1}{\hat{N}} \sum_{i \in S} \frac{1}{\pi_i} I\{y_i < y\}.$$

In this section, we compare via simulation the proposed estimator $\hat{G}_\pi$ in Equation (6) with Equations (13) and (14). We also compare their variance estimators and coverages of their 95% confidence intervals. Our simulation study will show that the proposed estimator in Equation (6) can be less biased than Equations (13) and (14). The observed coverages of the empirical likelihood confidence interval are also closer to the nominal value.

We generated $N = 10{,}000$ population values y_i from different distributions as in Davidson (2009), Qin et al. (2010) and Peng (2011), namely the χ^2, exponential, lognormal, Pareto and Weibull distributions. The different values of G_0 defined by Equation (2) are given in Table 1. We selected 2,000 randomized systematic samples of size $n = 200$ and 500. The inclusion probabilities π_i are generated from a linear model with y_i as covariate, in order to obtain a correlation of 0.7 between π_i and y_i. We chose this correlation to highlight the effect of the design. We use $B = 1{,}000$ replicates for the bootstrap procedures.

In Table 1, we have the observed relative bias (RB) and mean squared error (MSE) given by

$$RB(\hat{G}) := \frac{\hat{E}(\hat{G}) - G_0}{G_0} \times 100\%,$$

$$MSE(\hat{G}) := \hat{E}\{(\hat{G} - G_0)^2\}$$

for $\hat{G} = \hat{G}_\pi, \hat{G}_\pi^{(1)}$ and $\hat{G}_\pi^{(2)}$. Here, $\hat{E}(\cdot)$ denotes the means over the 2,000 observed values. The RB of $\hat{G}_\pi$ is slightly smaller than with $\hat{G}_\pi^{(2)}$. The RB of $\hat{G}_\pi^{(1)}$ tends to be the smallest for large values of G_0. However, $\hat{G}_\pi^{(1)}$ has the largest RB when G_0 is small. The MSE of $\hat{G}_\pi$ and $\hat{G}_\pi^{(2)}$ are similar. The MSE of $\hat{G}_\pi^{(1)}$ is slightly larger when $n = 200$. With $n = 500$, all the MSE are similar. From Table 1, we conclude that $\hat{G}_\pi$ tends to have the smallest bias with a MSE comparable to one observed for $\hat{G}_\pi^{(2)}$.

Table 1. Relative bias (%) and mean squared error (MSE) of $\hat{G}_\pi$, $\hat{G}_\pi^{(1)}$ and $\hat{G}_\pi^{(2)}$ for several distributions. G_0 is given by (3). The rows are sorted according to G_0.

			Relative bias (%)			MSE × 10,000		
n	Distributions	G_0	$\hat{G}_\pi$	$\hat{G}_\pi^{(1)}$	$\hat{G}_\pi^{(2)}$	$\hat{G}_\pi$	$\hat{G}_\pi^{(1)}$	$\hat{G}_\pi^{(2)}$
200	Pareto($\alpha = 10$, $\beta = 1$)	0.05	−0.4	8.5	−0.9	0.2	0.4	0.2
	Weibull($\alpha = 10$, $\beta = 1$)	0.07	−0.7	6.3	−1.2	0.2	0.3	0.2
	Pareto($\alpha = 5$, $\beta = 1$)	0.11	1.0	5.1	0.6	0.8	1.1	0.8
	Weibull($\alpha = 5$, $\beta = 1$)	0.13	−0.6	2.8	−1.0	0.6	0.7	0.6
	$\Gamma(\alpha = 10, \beta = 1)$	0.18	−0.9	2.0	−1.3	1.2	1.5	1.2
	$\Gamma(\alpha = 5, \beta = 1)$	0.25	−2.0	0.2	−2.4	2.4	2.6	2.5
	LogN($\mu = 0$, $\sigma = 0.5$)	0.28	−0.6	0.9	−1.0	1.8	1.8	1.8
	Exp($\lambda = 1$)	0.50	−0.9	−0.3	−1.4	6.2	6.0	6.4
	χ^2_1	0.64	−1.9	−1.4	−2.3	8.8	8.1	9.7
	$\Gamma(\alpha = 0.2, \beta = 1)$	0.80	0.0	0.1	−0.3	3.5	3.5	3.6
500	Pareto($\alpha = 10$, $\beta = 1$)	0.05	−0.2	3.4	−0.3	0.1	0.1	0.1
	Weibull($\alpha = 10$, $\beta = 1$)	0.07	−0.7	2.1	−0.9	0.1	0.1	0.1
	Pareto($\alpha = 5$, $\beta = 1$)	0.11	0.9	2.6	0.8	0.3	0.4	0.3
	Weibull($\alpha = 5$, $\beta = 1$)	0.13	−0.5	0.9	−0.7	0.2	0.2	0.2
	$\Gamma(\alpha = 10, \beta = 1)$	0.18	−0.5	0.7	−0.7	0.5	0.5	0.5
	$\Gamma(\alpha = 5, \beta = 1)$	0.25	−1.7	−0.9	−1.9	1.1	1.0	1.1
	LogN($\mu = 0$, $\sigma = 0.5$)	0.28	−0.4	0.2	−0.6	0.7	0.7	0.7
	Exp($\lambda = 1$)	0.50	−0.9	−0.7	−1.1	2.5	2.4	2.6
	χ^2_1	0.64	−1.7	−1.5	−1.9	4.6	4.3	4.9
	$\Gamma(\alpha = 0.2, \beta = 1)$	0.80	−0.1	0.0	−0.2	1.4	1.4	1.4

In Table 2, we have the observed coverages of the 95% confidence intervals. For $\hat{G}_\pi$, we consider two confidence intervals: The "*bootstrap confidence interval*" based on the 2.5% and 97.5% quantiles of the bootstrap (column "*Boot*"), and the empirical likelihood confidence intervals in Equation (12) (column "*EL*"). The usual confidence intervals based on linearised variance estimates are used for $\hat{G}_\pi^{(1)}$ and $\hat{G}_\pi^{(2)}$. The quantity G_0 is the target parameter on which the confidence intervals are based. The relative bias of the variance estimator

$$RB\{\hat{V}(\hat{G})\} := \frac{\hat{E}\{\hat{V}(\hat{G})\} - V(\hat{G})}{V(\hat{G})} \times 100\%$$

is given in the last three columns, where $V(\hat{G})$ denotes the observed variance. The bootstrap variance is used for $\hat{G}_\pi$. For $\hat{G}_\pi^{(1)}$ and $\hat{G}_\pi^{(2)}$, we use the linearisation variance estimates (e.g., Berger 2008; Langel and Tillé 2013) based on Hartley and Rao's (1962) variance estimator.

The observed coverages of the empirical likelihood approach are usually not significantly different from 95%, when the other coverages are different from 95%. The low coverages of $\hat{G}_\pi^{(1)}$ and $\hat{G}_\pi^{(2)}$ can be explained by lack of normality. With small values of G_0, the lower bounds of $\hat{G}_\pi^{(1)}$ and $\hat{G}_\pi^{(2)}$ can be negative. This could also explain the low coverage of $\hat{G}_\pi^{(1)}$. When the coverage of the empirical likelihood approach is significantly

Table 2. Observed coverages (%) of 95% confidence intervals of $\hat{G}_\pi$ (bootstrap and empirical likelihood), $\hat{G}_\pi^{(1)}$ and $\hat{G}_\pi^{(2)}$. Relative bias RB$\{\hat{V}(\hat{G})\}$ *(%), of the bootstrap variance estimator of $\hat{G}_\pi$ and the linearised variance of $\hat{G}_\pi^{(1)}$ and $\hat{G}_\pi^{(2)}$. Several distributions are considered. The rows are sorted according to G_0 (see Table 1 for the values of G_0).*

		Coverages (%)						
		$\hat{G}_\pi$				RB$\{\hat{V}(\hat{G})\}$ (%)		
n	Distributions	Boot	(11)	$\hat{G}_\pi^{(1)}$	$\hat{G}_\pi^{(2)}$	$\hat{G}_\pi$	$\hat{G}_\pi^{(1)}$	$\hat{G}_\pi^{(2)}$
200	Pareto($\alpha = 10$, $\beta = 1$)	94.2	94.6	83.2†	94.0†	2.6	1.6	1.6
	Weibull($\alpha = 10$, $\beta = 1$)	93.5†	94.6	85.6†	93.4†	4.8	3.7	4.0
	Pareto($\alpha = 5$, $\beta = 1$)	94.4	95.1	92.6†	94.3	4.4	2.1	2.2
	Weibull($\alpha = 5$, $\beta = 1$)	93.5†	94.5	93.3†	93.3†	3.9	2.7	3.2
	$\Gamma(\alpha = 10, \beta = 1)$	91.1†	92.9†	94.6	90.6†	−7.6	−17.5	−5.5
	$\Gamma(\alpha = 5, \beta = 1)$	89.4†	92.7†	93.4†	89.1†	−10.9	−2.6	−9.9
	LogN($\mu = 0$, $\sigma = 0.5$)	93.7†	95.4	94.8	93.7†	−1.0	−1.7	−2.0
	Exp($\lambda = 1$)	90.7†	92.7†	92.7†	90.4†	−13.7	−14.9	−14.4
	χ^2_1	89.1†	90.4†	90.8†	89.3†	−19.7	−20.8	−18.8
	$\Gamma(\alpha = 0.2, \beta = 1)$	94.7	95.4	94.4	94.6	−5.7	0.0	−2.8
500	Pareto($\alpha = 10$, $\beta = 1$)	95.1	95.2	90.1†	94.3	4.9	−0.9	−0.9
	Weibull($\alpha = 10$, $\beta = 1$)	93.5†	94.6	85.6†	93.4†	4.8	3.7	4.0
	Pareto($\alpha = 5$, $\beta = 1$)	95.1	94.8	93.1†	94.3	4.2	−3.6	−3.9
	Weibull($\alpha = 5$, $\beta = 1$)	93.2†	95.1	94.4	93.0†	0.6	−3.7	−3.5
	$\Gamma(\alpha = 10, \beta = 1)$	93.4†	94.7	94.4	93.2†	1.5	−2.7	−0.5
	$\Gamma(\alpha = 5, \beta = 1)$	88.6†	92.9†	92.3†	88.3†	−1.2	2.2	−1.9
	LogN($\mu = 0$, $\sigma = 0.5$)	94.4†	95.4	94.6	94.2	−1.0	−1.7	−2.0
	Exp($\lambda = 1$)	93.1†	93.8†	93.7†	92.8†	−6.0	−7.7	−8.2
	χ^2_1	87.7†	87.5†	89.0†	87.6†	3.0	−1.6	0.6
	$\Gamma(\alpha = 0.2, \beta = 1)$	95.5	96.0	94.6	95.2	8.6	0.6	1.6

† Coverage rates significantly different from 95%: p-value $\leq$ 0.05.

different from 95%, the other coverages are also significantly different (distributions $\Gamma(\alpha = 5, \beta = 1)$, $Exp(\lambda = 1)$ and χ^2_1). The distribution $\Gamma(\alpha = 10, \beta = 1)$ is an exception, because $\hat{G}_\pi^{(1)}$ has the best coverage, but with a biased variance estimator. We have observed one sample of size $n = 200$ with a negative lower bound for the confidence interval of Equation (13). This occurs with the data generated from a χ^2-distribution.

The RB of the variance of $\hat{G}_\pi$ can be large with $n = 200$, because they are based on bootstrap. However, with $n = 500$, all the RB are similar, and $\hat{G}_\pi$ may have the smallest RB. When $n = 200$, we have larger RB for large values of G_0 (distributions $Exp(\lambda = 1)$ and χ^2_1 and $\Gamma(\alpha = 0.2, \beta = 1)$).

In Table 3, we have the observed average length of the 95% confidence intervals, as well as the observed "*coefficient of variation*" (CV) of the lengths. The average length is very similar and in line with the coverages observed in Table 2, because confidence intervals with large coverage tend to be larger on average.

A small CV implies more stable confidence intervals, but this does not imply observed coverages closer to 95%. The CV of the bootstrap confidence intervals tends to be the smallest, but with observed coverage significantly different from 95%. For the Pareto and Weibull distribution, the CV of Equation (11) is slightly larger than the other confidence intervals, which have coverages usually different from 95%. This effect is more

Table 3. Observed Average Length of 95% confidence intervals and observed coeffcient of the lengths variation (CV) in percent. The rows are sorted according to G_0 (see Table 1 for the values of G_0).

		Average Lengths				CV(Lengths) %			
		$\hat{G}_\pi$				$\hat{G}_\pi$			
n	Distributions	Boot	(11)	$\hat{G}_\pi^{(1)}$	$\hat{G}_\pi^{(2)}$	Boot	(11)	$\hat{G}_\pi^{(1)}$	$\hat{G}_\pi^{(2)}$
200	Pareto($\alpha = 10, \beta = 1$)	0.017	0.018	0.017	0.017	16.4	18.1	16.2	16.2
	Weibull($\alpha = 10, \beta = 1$)	0.016	0.016	0.016	0.016	13.5	13.3	13.3	13.3
	Pareto($\alpha = 5, \beta = 1$)	0.036	0.037	0.035	0.035	14.6	15.1	14.2	14.2
	Weibull($\alpha = 5, \beta = 1$)	0.030	0.030	0.029	0.030	11.8	11.7	11.5	11.5
	$\Gamma(\alpha = 10, \beta = 1)$	0.039	0.040	0.041	0.039	23.1	38.4	66.9	28.5
	$\Gamma(\alpha = 5, \beta = 1)$	0.054	0.056	0.057	0.054	23.7	37.8	43.8	28.8
	LogN($\mu = 0, \sigma = 0.5$)	0.052	0.053	0.052	0.052	11.9	12.7	15.9	13.3
	Exp($\lambda = 1$)	0.089	0.092	0.089	0.090	19.3	30.9	26.2	23.7
	χ^2_1	0.089	0.095	0.091	0.091	27.4	38.5	40.1	33.8
	$\Gamma(\alpha = 0.2, \beta = 1)$	0.076	0.077	0.074	0.076	9.0	9.5	8.1	9.0
500	Pareto($\alpha = 10, \beta = 1$)	0.011	0.011	0.011	0.011	11.0	11.1	10.4	10.4
	Weibull($\alpha = 10, \beta = 1$)	0.010	0.010	0.010	0.010	8.6	8.6	8.3	8.3
	Pareto($\alpha = 5, \beta = 1$)	0.023	0.023	0.022	0.022	9.3	9.4	8.6	8.6
	Weibull($\alpha = 5, \beta = 1$)	0.019	0.019	0.019	0.019	8.2	8.1	7.6	7.6
	$\Gamma(\alpha = 10, \beta = 1)$	0.025	0.026	0.026	0.025	20.9	47.5	79.4	23.9
	$\Gamma(\alpha = 5, \beta = 1)$	0.036	0.037	0.036	0.035	22.0	37.6	33.9	26.8
	LogN($\mu = 0, \sigma = 0.5$)	0.033	0.033	0.032	0.032	9.9	9.9	12.0	10.9
	Exp($\lambda = 1$)	0.058	0.060	0.058	0.058	20.1	34.9	27.5	23.9
	χ^2_1	0.060	0.064	0.060	0.061	32.3	52.4	49.5	39.1
	$\Gamma(\alpha = 0.2, \beta = 1)$	0.048	0.048	0.046	0.046	5.7	6.0	5.1	5.1

pronounced with $n = 200$. With the Gamma and χ^2-distributions, we have a small CV with bootstrap and $\hat{G}_{\pi}^{(2)}$, but with very low coverages.

6. Discussion

Our simulation study shows the proposed estimator is usually less biased than the customary estimators. The observed coverages of the empirical likelihood confidence interval are also closer to the nominal value. We considered single stage design. However, the proposed approach can be extended for multi-stage designs with unit nonresponse, using Berger's (2018a) approach combined with bootstrap. Auxiliary information has not been considered for simplicity. Calibration weights can be used within Equation (6). The proposed empirical likelihood approach can also take into account some auxiliary information, by adding additional constraints within Equation (10) (see Berger and Torres 2016; Berger 2018a,b, for more details). These additional constraints imply that $\hat{G}_{EL}$ will be different but usually close to $\hat{G}_{\pi}$, because $\hat{G}_{EL}$ is based on calibrated weights.

7. References

Beach, C.M. and R. Davidson. 1983. "Distribution-free statistical inference with Lorenz curves and income shares." *The Review of Economic Studies* 50(4): 723–735. DOI: https://doi.org/10.2307/2297772.

Berger, Y.G. 2008. "A note on the asymptotic equivalence of jackknife and linearization variance estimation for the Gini coefficient." *Journal of Official Statistics* 24-4: 541–555. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/a-note-on-the-asymptotic-equivalence-of-jackknife-and-linearization-variance-estimation-for-the-gini-coefficient.pdf (accessed April 2020).

Berger, Y.G. 2018a. "An empirical likelihood approach under cluster sampling with missing observations." *Annals of the Institute of Statistical Mathematics*. DOI: https://doi.org/10.1007/s10463-018-0681-x.

Berger, Y.G. 2018b. "Empirical likelihood approaches in survey sampling." *The Survey Statistician* 78: 22–31. Available at: http://isi-iass.org/home/wp-content/uploads/N78-2018-07-ISSN.pdf (accessed April 2020).

Berger, Y.G. and O.D.L.R. Torres. 2012. "A unified theory of empirical likelihood ratio confidence intervals for survey data with unequal probabilities." Proceedings of the Survey Research Method Section of the American Statistical Association, Joint Statistical Meetings, San Diego: 15. Available at: http://www.asasrms.org/Proceedings (accessed November 2019).

Berger, Y.G. and O.D.L.R. Torres. 2014. "Empirical likelihood confidence intervals: an application to the EU-SILC household surveys." In *Contribution to Sampling Statistics, Contribution to Statistics*, edited by F. Mecatti, P.L. Conti, and M.G. Ranalli, 65–84. Springer.

Berger, Y.G. and O.D.L.R. Torres. 2016. "An empirical likelihood approach for inference under complex sampling design." *Journal of the Royal Statistical Society Series B* 78(2): 319–341. DOI: https://doi.org/10.1111/rssb.12115.

Bhattacharya, D. 2007. "Inference on inequality from household survey data." *Journal of Econometrics* 137(2): 674–707. DOI: https://doi.org/10.1016/j.jeconom.2005.09.003.

Binder, D.A. and M.S. Kovaćević. 1995. "Estimating some measure of income inequality from survey data: an application of the estimating equation approach." *Survey Methodology* 21(2): 137–145. Available at: https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X199500214396 (accessed April 2020).

Bishop, J., J.P. Formby, and B. Zheng. 1997. "Statistical inference and the Sen index of poverty." *International Economic Review* 381–387. DOI: https://doi.org/10.2307/2527379.

Brent, R.P. 1973. *Algorithms for Minimization without Derivatives*. New-Jersey: Prentice-Hall. ISBN 0-13-022335-2.

Davidson, R. 2009. "Reliable inference for the Gini index." *Journal of Econometrics* 150(1): 30–40. DOI: https://doi.org/10.1016/j.jeconom.2008.11.004.

Dekker, T.J. 1969. "Finding a zero by means of successive linear interpolation." In *Constructive Aspects of the Fundamental Theorem of Algebra*, edited by B. Dejon and P. Henrici, 37–489. Handbook of Statistics, London: Wiley-Interscience.

Gajdos, T. and J.A. Weymark. 2005. "Multidimensional generalized Gini indices." *Economic Theory* 26(3): 471–496. DOI: https://doi.org/10.1007/s00199-004-0529-x.

Gastwirth, J.L. and M. Gail. 1985. "Simple asymptotically distribution-free methods for comparing Lorenz curves and Gini indices obtained from complete data." *Advances in Econometrics* 4: 229–243. Available at: https://scholar.google.com/scholar_lookup?title=Simple%20Asymptotically%20Distribution-free%20Methods%20for%20Comparing%20Lorenz%20Curves%20and%20Gini%20Indices%20Obtained%20from%20Complete%20Data&author=JL.%20Gastwirth&author=MH.%20Gail&pages=229-243&publication_year=1985 (accessed April 2020).

Giles, D. 2004. "Calculating a standard error for the Gini coefficient: some further results." *Oxford Bulletin of Economics and Statistics* 66(3): 425–433. DOI: https://doi.org/10.1111/j.1468-0084.2004.00086.x.

Gini, C. 1914. "Sulla Misura Della Concentrazione e Cella Variabilitá dei Caratteri." *Atti del Reale Istituto veneto di scienze, lettere ed arti*, 73: 1203–1248. Available at: https://EconPapers.repec.org/RePEc:mtn:ancoec:0501 (accessed April 2020).

Giorgi, G.M. and C. Gigliarano. 2017. "The Gini concentration index: a review of the inference literature." *Journal of Economic Surveys* 31(4): 1130–1148. DOI: https://doi.org/10.1111/joes.12185.

Glasser, G.J. 1962. "Variance formulas for the mean difference and coefficient of concentration." *Journal of the American Statistical Association* 57(299): 648–654. DOI: https://doi.org/10.1080/01621459.1962.10500553.

Hájek, J. 1971. "Comment on a paper by Basu, D." In *Foundations of Statistical Inference*, edited by V.P. Godambe and D.A. Sprott, 236. Toronto: Holt, Rinehart & Winston.

Hartley, H.O. and J.N.K. Rao. 1962. "Sampling with unequal probabilities without replacement." *The Annals of Mathematical Statistics* 33: 350–374. DOI: https://doi.org/10.1214/aoms/1177704564.

Hoeffding, W. 1948. "A non-parametric test of independence." *The annals of Mathematical Statistics*, 546–557. DOI: https://doi.org/10.1214/aoms/1177730150.

Karagiannis, E. and M. Kovaćević. 2000. "A method to calculate the Jack-250 knife variance estimator for the Gini coefficient." *Oxford Bulletin of Economics and Statistics* 62(1): 119–122. DOI: https://doi.org/10.1111/1468-0084.00163.

Koshevoy, G. and K. Mosler. 1997. "Multivariate Gini indices." *Journal of Multivariate Analysis* 60(2): 252–276. DOI: https://doi.org/10.1006/jmva.1996.1655.

Langel, M. and Y. Tillé. 2013. "Variance estimation of the Gini index: revisiting a result several times published." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 176(2): 521–540. DOI: https://doi.org/10.1111/j.1467-985X.2012.01048.x.

Lorenz, M.O. 1905. "Methods of measuring the concentration of wealth." *Publications of the American Statistical Association* 9(70): 209–219. DOI: https://doi.org/10.2307/2276207.

Mills, J.A. and S. Zandvakili. 1997. "Statistical inference via bootstrapping for measures of inequality." *Journal of Applied Econometrics* 12(2): 133–150. DOI: https://doi.org/10.1002/(SICI)1099-1255(199703)12:2133:AID-JAE4333.0.CO;2-H.

Modarres, R. and J.L. Gastwirth. 2006. "A cautionary note on estimating the standard error of the Gini index of inequality." *Oxford Bulletin of Economics and Statistics* 68(3): 385–390. DOI: https://doi.org/10.1111/j.1468-0084.2006.00167.x.

Nair, U.S. 1936. "The standard error of Gini's mean difference." *Biometrika* 28(3/4): 428–436. DOI: https://doi.org/10.1093/biomet/28.3-4.428.

Neyman, J. 1938. "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection." *Journal of the Royal Statistical Society* 97(4): 558–625. DOI: https://doi.org/10.2307/2342192.

Nygård, F. and A. Sandström. 1989. "Income inequality measures based on sample surveys." *Journal of Econometrics* 42(1): 81–95. DOI: https://doi.org/10.1016/0304-4076(89)90077-8.

Ogwang, T. 2000. "A convenient method of computing the Gini index and its standard error." *Oxford Bulletin of Economics and Statistics* 62(1): 123–129. DOI: https://doi.org/10.1111/1468-0084.00164.

Owen, A.B. 1988. "Empirical Likelihood Ratio Confidence Intervals for a Single Functional." *Biometrika* 75(2): 237–249. DOI: https://doi.org/10.1093/biomet/75.2.237.

Peng, L. 2011. "Empirical likelihood methods for the Gini index." *Australian & New Zealand Journal of Statistics* 53(2): 131–139. DOI: https://doi.org/10.1111/j.1467-842X.2011.00614.x.

Qin, Y., J.N.K. Rao, and C. Wu. 2010. "Empirical likelihood confidence intervals for the Gini measure of income inequality." *Economic Modelling* 27: 1429–1435. DOI: https://doi.org/10.1016/j.econmod.2010.07.015.

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. Available at: https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X199200214486 (accessed April 2020).

Rust, K.F. and J.N.K. Rao. 1996. "Variance estimation for complex surveys using replication techniques." *Biometrika* 5(3): 281310. DOI: https://doi.org/10.1177/096228029600500305.

Sandström, A., B. Waldén, and J.H. Wretman. 1985. *Variance estimators of the Gini coefficient: simple random sampling*. Sweden. Available at: https://www.scb.se/contentassets/7004f1a7effe4e8690548165695b1864/pm-p-1985-17-green.pdf (accessed April 2020).

Sandström, A., J.H. Wretman, and B. Waldén. 1988. "Variance estimators of the Gini coefficient – probability sampling." *Journal of Business & Economic Statistics* 6(1): 113–119. DOI: https://doi.org/10.1080/07350015.1988.10509643.

Schezhtman, E. and S. Yitzhaki. 1987. "A Measure of Association Based on Gini's Mean Difference." *Communications in Statistics-Theory and Methods* 16(1): 207–231. DOI: https://doi.org/10.1080/03610928708829359.

Sendler, W. 1979. "On statistical inference in concentration measurement." *Metrika* 26(1): 109–122. DOI: https://doi.org/10.1007/BF01893478.

Shao, J. 1994. "L-Statistics in Complex Survey Problems." *The Annals of Statistics* 22(2): 946–967. DOI: https://doi.org/10.1214/aos/1176325505.

Wu, C. and J.N.K. Rao. 2006. "Pseudo-empirical likelihood ratio confidence intervals for complex surveys." *Canadian Journal of Statistics* 34(3): 359–375. DOI: https://doi.org/10.1002/cjs.5550340301.

Yitzhaki, S. 1991. "Calculating jackknife variance estimators for parameters of the Gini method." *Journal of Business & Economic Statistics* 9(2): 235–239. DOI: https://doi.org/10.1080/07350015.1991.10509849.

Yitzhaki, S. 1998. "More than a dozen alternative ways of spelling Gini." *Research on Economic Inequality* 8: 13–30. DOI: https://doi.org/10.1.1.365.4196.

Received October 2018
Revised March 2019
Accepted December 2019

sciendo

Journal of Official Statistics, Vol. 36, No. 2, 2020, pp. 251–274, http://dx.doi.org/10.2478/JOS-2020-0014

Estimating Literacy Levels at a Detailed Regional Level: an Application Using Dutch Data

Ineke Bijlsma[1], *Jan van den Brakel*[1], *Rolf van der Velden*[1], *and Jim Allen*[1]

Policy measures to combat low literacy are often targeted at municipalities or regions with low levels of literacy. However, current surveys on literacy do not contain enough observations at this level to allow for reliable estimates when using only direct estimation techniques. To provide more reliable results at a detailed regional level, alternative methods must be used.

The aim of this article is to obtain literacy estimates at the municipality level using model-based small area estimation techniques in a hierarchical Bayesian framework. To do so, we link Dutch Labour Force Survey data to the most recent literacy survey available, that of the Programme for the International Assessment of Adult Competencies (PIAAC). We estimate the average literacy score, as well as the percentage of people with a low literacy level. Variance estimators for our small area predictions explicitly account for the imputation uncertainty in the PIAAC estimates. The proposed estimation method improves the precision of the area estimates, making it possible to break down the national figures by municipality.

Key words: Literacy; basic skills; municipality; region; small area estimation.

1. Introduction

Research shows that cognitive skills play an important role in individual life chances (Coulombe and Tremblay 2007; Hanushek and Woessmann 2008, 2011). People with high skill proficiency levels earn more, are more often employed, and generally face fewer economic disadvantages. Moreover, they are more often engaged in civic and social activities (Organisation for Economic Co-operation and Development (OECD) 2013a).

Generally, the skill levels in the Netherlands are among the highest in the world. In the Programme for the International Assessment of Adult Competencies (PIAAC) of 2012, the Netherlands ranked third in literacy, just behind Japan and Finland. Even so, there are still around 1.3 million people (11.9%) in the population of 16- to 65-year-olds who do not have the literacy skills necessary to function well in society (Buisman et al. 2013). The cost

[1] Maastricht University, ROA, P.O. Box 616, 6200 MD Maastricht, The Netherlands. Emails: i.bijlsma@maastrichtuniversity.nl, j.vandenbrakel@maastrichtuniversity.nl, r.vandervelden@maastrichtuniversity.nl, and j.allen@maastrichtuniversity.nl

Acknowledgments: The views expressed in this article are those of the authors and do not necessarily reflect the policy of Statistics Netherlands. This study was made possible by a grant from the Reading and Writing Foundation, which other than funding had no role in this research. The authors are grateful to the unknown referees and the associate editor for reading and commenting on a former draft of this article. We also thank Olivier Marie, Stefa Hirsch, participants of the SAE 2016 conference (17–19 August 2016, Maastricht, The Netherlands) and the third PIAAC International Conference (6–8 November 2016, Madrid, Spain) for useful comments and suggestions.

of low literacy in the Netherlands is estimated to be some 550 million euros per year (PriceWaterhouseCoopers 2013).

As policy aimed at increasing literacy levels is often decentralized, local and regional governments need reliable data on the literacy levels in their particular municipality or region. However, this is usually not available, since most literacy surveys such as PIAAC focus on the national level. To illustrate the problem: The Dutch PIAAC sample contains about 5,000 observations. However, the Netherlands comprises 415 municipalities, and only the four biggest cities in the Netherlands have more than 90 observations in the PIAAC sample, while roughly half of the municipalities have fewer than 20 observations. The use of direct estimators would result in unacceptably large design variances. To increase the precision of municipal estimates, model-based small area estimation (SAE) techniques are applied in this article. These methods assume an explicit statistical model to increase the effective sample size of each separate area.

The basic idea of this regression method is that we assume that our dependent variable, literacy, is closely linked to personal characteristics such as age, gender, education, and labor status, which are also available in large auxiliary data sets. We also make the necessary assumption that the way these characteristics are linked is similar at both the national and detailed regional levels. Therefore, with detailed information for these characteristics at the regional level, it is possible to make more accurate model-based literacy predictions per municipality: a synthetic estimate. Unexplained variation between the areas is modeled with a random component in a multilevel model.

Model-based small area predictors can be expressed as the weighted average between the direct estimates based on PIAAC data and the aforementioned synthetic estimates, where the weights are based on the accuracy measures of the two estimators. If the underlying assumptions hold, this allows us to greatly reduce the variance of the estimates while introducing only limited bias to the estimates.

SAE techniques are widely applied in social and economic sciences to produce reliable statistical information in detailed breakdowns. Taylor et al. (2016) use synthetic estimates to predict expected levels of limiting long-term illnesses. The World Bank (2002) applies a synthetic estimation procedure proposed by Elbers et al. (2003) to estimate poverty and income inequality in developing countries. The U.S. Census Bureau applies an SAE approach based on the Fay and Herriot (1979) model to estimate income at low regional levels. These estimates are used to determine fund allocations to local government units. The National Research Council (2000) also used the method of Fay-Herriot to produce county estimates of poor school-aged children in the United States for the allocation of supporting funds. Statistics Netherlands applies time series SAE methods to calculate official monthly unemployment Figures (Van den Brakel and Krieg 2015). Finally, Tighe et al. (2010) applied hierarchical Bayesian models to obtain reliable estimates for low-incidence groups defined by religion or ethnicity not included in the U.S. Census Bureau.

To the best of our knowledge, SAE techniques in the context of literacy skills have only been applied sparsely, and take a quite different approach than the one we present here. Schmid et al. (2017) use self-assessed literacy from the Demographic and Health Survey in combination with mobile phone data to estimate literacy in Senegal, as a way to use alternative data sources instead of requiring statistics on socio-demographic indicators. Gibson and Hewson (2012) use UK census data and SAE modeling to obtain synthetic

estimates of literacy levels in detailed geographical areas. Yamamoto (2014) adopts a similar approach to produce synthetic estimates for the different Canadian provinces.

While these two papers focus on synthetic estimates only, the contribution of this article is the application of SAE techniques to estimate municipalities' literacy levels that are a weighted average of direct and synthetic estimates, with the weights based on the uncertainty measures of both estimates. This approach has the advantage that, in large municipalities with relatively large sample sizes, the direct estimates make a relatively large contribution to the final estimate, whereas in small municipalities, the final estimate is dominated by the synthetic estimator. The PIAAC data setup presents a number of challenges that prevent straightforward estimations. Addressing these challenges is novel in the application of SAE techniques. Respondents were randomly assigned to (parts of) the literacy tests. This requires imputation techniques to account for missing observations. Moreover, the PIAAC tests follow an adaptive design, so that respondents are assigned items that are close to their expected proficiency levels, based on the scores of previous questions. The model follows an item response theory (IRT) approach, which assumes that the scores on the tests are based on a latent construct that cannot be measured directly. Instead, for each respondent, ten plausible values are calculated and several replicate weights are constructed, which can be seen as a form of multiple imputation. This approach allows for the construction of point estimates as well as variance estimates for literacy. We use both a unit-level model (Battese et al. 1988) and an area-level model (Fay and Herriot 1979) and detail how to incorporate this structure into our SAE approach.

Our article is organized as follows. Section 2 covers the definition of literacy, as well as the data description. Section 3 details the techniques of the small area predictors for this application. Section 4 presents the selected models and their fit. Section 5 evaluates the model and presents robustness checks. Section 6 reports the results of our analysis and Section 7 concludes the article.

2. Definition of Literacy and Data Description

2.1. PIAAC – Primary Data Source

The data set we are using is the 2012 PIAAC survey. It is designed to map skills and competencies in developed countries, measuring the numeracy, literacy, and problem solving skills of adults. In addition, it collects a range of information on how often respondents use these skills.

Literacy in PIAAC is defined as "the ability to understand, evaluate, use and engage with written texts to participate in society, to achieve one's goals, and develop one's knowledge and potential" (OECD 2013a, 59). It does not include the ability to write or produce texts, but focuses on the ability of an individual to interact with written text. It is this definition that will be used throughout the article.

Data collection in the Netherlands took place from August 1, 2011, to March 31, 2012, and was undertaken in the respondents' homes. The target population was between 16 and 65 years of age, residing in the country at the time the data were collected. For the Netherlands, 5,170 respondents were randomly selected by one-stage stratified simple random sampling without replacement from the Dutch population register. Strata were formed by

municipalities. The sample weights are based on the sampling design. The response rate, as defined by complete cases divided by eligible cases, was 51% (OECD 2013b).

The PIAAC survey used specific data collection modes and procedures to measure skill proficiency levels (for details, see OECD 2013c). For the literacy domain, the questions differed in content, cognitive strategies, and context. A multistage adaptive design was used between the items and an algorithm determined the next item depending on the responses. This survey design was such that different groups of respondents were routed to items with potentially various degrees of difficulty, disallowing direct comparisons between the respondents' test scores. Therefore, the item responses were first fitted to an IRT model. After item calibration, the IRT model was combined with a latent regression model using information from the background questionnaire in a population model to further improve accuracy. From this step, 10 plausible values were drawn on a scale from zero to 500. Lastly, a replication approach (Johnson and Rust 1992) was used to estimate the sampling variability as well as the imputation variance associated with the plausible values. The percentage of respondents in the Netherlands who were unable to complete the questionnaire due to literacy-related issues is 2.3%; no proficiency scores were estimated for this group, but they were included in the weighting (OECD 2013b). The effect of list-wise deletion of these cases is therefore limited.

Variance estimation, taking into account the sample design, the selection process, the weighting adjustment, and the measurement error through imputation, is carried out using a replication approach. For the Netherlands, a paired jackknife estimator was used with 80 replicate weights. To take this survey design into account, we used the Stata module PIAACTOOLS of Pokropek and Jakubowski (2013). A detailed description of the construction of the variance term, as well as the above imputation, can be found in OECD (2013c).

Literacy scores are categorized at multiple levels based on the scoring range. Level 1 literacy starts at a score 176, and every 50 points above indicates an additional level, up to Level 5 (376 points or higher). At Level 1 (range 176–225), one can complete simple forms, understand basic vocabulary, and read continuous texts, but would have trouble making low-level inferences. For reference, Level 3 requires multiple steps to access the correct information and at Level 5 one can work with multiple, dense texts and conflicting information. These levels are described in full in OECD (2013b).

One straightforward method for describing the literacy levels in a region would be to look at the average test score for literacy. This is a good way of providing a quick snapshot of the literacy level. A limitation, however, is that it provides no further information as to how literacy levels are distributed within regions. Another measure would be to look at the proportion of low literates per area. We define someone as *low literate* when that individual has literacy Level 1 or below. This measure would be most important for policy making, as this group would benefit the most from policy interventions. A disadvantage of this measure is that information is lost due to its dichotomous nature. Taken together, both measures – the average score and the proportion of low literates – provide the best picture of the situation concerning literacy levels in a region.

The total number of respondents in PIAAC is 5,170, but for some respondents the municipality is unknown. We are left with 5,073 respondents, whose statistics are given below (see Table 1). The average score across respondents is in the lower half of Level 3 (276–325), with only about 12% at Level 1 or below (225 or below).

Table 1. Summary of the statistics of the target sample (PIAAC).

	Mean	St. Error	Lower Bound	Upper Bound
Average Score	283.94	0.68	282.61	285.27
% Low Literates	12.00	0.46	11.07	12.86

In Section 3, two different small area estimation models are applied. The area level model (Fay and Herriot 1979) use direct estimates for the target variable and their variances at the level of the areas as input for the model. The unit level model (Battese et al. 1988) use the observations of the sampling units as input for the model. Both models are multilevel models and need auxiliary information for the fixed effect part of the model. The area level model can only use auxiliary information that is aggregated at the level of the area (municipality). The unit level model can use both auxiliary information at the level of the sampling units (individuals) and auxiliary information aggregated at the level of the areas. As stated in the introduction, we are interested in both the average literacy score and the percentage of low literacy per municipality. We estimate the literacy score using the unit-level model and low literacy using the area-level model (dichotomous); we expand on the construction of the dependent variables under *Literacy Measures*.

2.2. *Labor Force Survey (LFS) – Data Source for Auxiliary Information*

SAE requires auxiliary data that include personal characteristics that are closely linked to literacy levels. The Dutch LFS's features (large sample sizes, good overlap in questions about personal characteristics) make it a good choice for auxiliary data.

In our selected timeframe, interviews for the LFS took place face to face and by phone. The weights are calculated in two steps using general regression estimators (Särndal et al. 1992). In the first step, design weights are derived from the sample design and account for differences in selection probabilities. In a second step, the design weights are calibrated to available auxiliary information for which the true population distributions are known from registrations to correct, at least partially, for selective nonresponse.

To ensure sufficient data from each area, we chose to include three years of LFS data: 2010, 2011 and 2012, that is, years close to the data collection period for PIAAC. We apply the same age restriction (between 16 and 65 years old) as in the PIAAC survey.

The LFS is based on a household sample. All household members aged 15 years and older are observed. When a household member cannot be contacted, proxy interviewing is allowed by members of the same household. Households in which one or more of the selected persons do not respond for themselves or in a proxy interview are treated as non-responding households.

The total response and nonresponse numbers can be found in the *Methods and definitions* of the LFS data (Statistics Netherlands 2010; 2011; 2012), with a minimum response of roughly 63% of the approached households. This results in about 41,000 completely responding households on a yearly basis, and thus about 123,000 over three years (with a maximum of eight persons per household).

Since the LFS has a rotating panel design, people were asked multiple times to participate and thus are included multiple times. We weight these people over the number

of samples within our selection, so that those who are covered multiple times in the data set are not oversampled. This leaves us with 309,000 unique respondents (with a rough average of 2.5 persons per household).

3. Small Area Estimation

Sample surveys are usually designed to meet minimum precision requirements for sample estimates at national level and at the level of planned domains using standard direct estimators. For other unplanned domains or subpopulations, the sample size is frequently too small to create reliable estimates based on direct estimators. Sample size is restricted by available resources and time and, in many surveys, it is too costly to sample a large number of individuals within each subpopulation of interest. In such cases, model-based inference methods from the literature on SAE can be considered as an alternative. SAE refers to estimation procedures that explicitly rely on a statistical model that increases the effective sample size of a particular domain with sample information from other domains (cross-sectional correlations) or preceding sampling periods (temporal correlations). The extent to which the precision of direct estimates is improved with these methods depends on the availability of auxiliary data contained in register data sets or large surveys, such as the LFS.

A large amount of SAE procedures are available in the literature. See Rao and Molina (2015) for a detailed overview, or Pfeffermann (2013) for a more summarized overview. In this article, we have chosen a multilevel modeling approach. The models are fitted in a *hierarchical Bayesian* (HB) framework. All models, including the model selection measures, were run using the fSAE function in the software program R, available via the hbsae package (Version 1.0, available in the Comprehensive R Archive Network; Boonstra 2015).

It is important to keep some things in mind when interpreting the results from SAE. In particular, model miss-specification can result in biased domain predictions. One important possible bias is due to the assumption that the relations between literacy and personal characteristics at the national level are the same at the regional level. While we do not expect the literacy model to have regional variation, violation of this assumption can lead to large differences between the regional estimations and the true regional literacy.

3.1. Literacy Measures

As stated earlier, we are interested in two measures of literacy per area: the average score and the percentage of low literates. In the first case, the dependent variable y is continuous per individual and area and we assume that y has a linear relation with the chosen covariates X. In this case, we use the basic unit-level model originally proposed by Battese et al. (1988), where the input variables for the model are individual measurements obtained from the sampling units. We go into more detail in the section below on the unit-level model.

In the second case regarding the percentage of low literates, the dependent variable is dichotomous at the individual level, since each plausible value will be binary, equal to one if the score is below the low-literacy cutoff point of 226 and zero otherwise. We decided to model the percentage of low literates with a basic area-level model, as originally proposed by Fay and Herriot (1979), as the *hbsae* package has no support for binary outcome

variables that would be necessary for a unit-level model. In the next two sections, we elaborate both the area-level model and the unit-level model. Afterwards, we explain how we incorporated the PIAAC imputation structure in the estimations.

3.2. Area-Level Model

The input for the area-level model is provided by the direct estimates for the areas. Let y_{ia} denote the average of the ten plausible values of an individual i who belongs to municipality a, as observed in the original survey data (PIAAC). Specific for the area-level model, we transform each y_{ia} in a dichotomous value, as described in the above paragraph.

Then, the average of these values is used to construct the area average of literacy, for example, $\bar{y}_a$, using the paired jackknife estimator (see also Section 2). The jackknife is used to estimate the variance of $\bar{y}_a$, denoted Ψ_a^2, and accounts for sampling error, the uncertainty of multiple imputation for missing values, and the uncertainty of the IRT model underlying the adaptive tests for literacy, using both replicate weights and plausible values. Therefore, it takes fully into account the uncertainty resulting from the PIAAC questionnaire design (OECD 2013c). Furthermore, let $\bar{\mathbf{X}}_{\boldsymbol{a}}$ denote the vector with the population means of the auxiliary variables derived from the LFS used for calibration. The sample area means for the auxiliary variables derived from the PIAAC sample are denoted $\bar{\mathbf{x}}_{\boldsymbol{a}}$. Survey errors regarding the estimation of $\bar{\mathbf{X}}_{\boldsymbol{a}}$ from the LFS are assumed to be small enough to be negligible and are not taken into account.

In a first step, direct estimates for the target variable for each area are obtained using the survey regression estimator $\hat{y}_a^{surv}$:

$$\hat{y}_a^{surv} = \bar{y}_a + (\bar{\mathbf{X}}_{\boldsymbol{a}} - \bar{\mathbf{x}}_{\boldsymbol{a}})^{\mathrm{t}}\boldsymbol{\beta},$$

where $\boldsymbol{\beta}$ is the vector with regression coefficients from the linear model that describes the relation between the target variable y and the auxiliary variables x. These direct estimates are the input for the area level or Fay–Herriot model:

$$\hat{y}_a^{surv} = \alpha + \bar{\mathbf{X}}_{\boldsymbol{a}}\boldsymbol{\beta} + \mathrm{u}_a + \mathrm{e}_a \tag{1}$$

where α is the intercept, $\bar{\mathbf{X}}_{\boldsymbol{a}}$ the area covariate averages, $\boldsymbol{\beta}$ the vector of coefficients of covariates, and u_a a random effect to take into account area-level variation not explained by the fixed part of the equation. The random effects are assumed to be normally and independently distributed, with an expected value equal to zero and model variance σ^2. Finally, e_a is an independently distributed sampling error that has expected value zero and sampling variance Ψ_a^2. Based on this model, the best linear unbiased predictor (BLUP) estimator for the area means is equal to (Rao and Molina 2015):

$$\hat{y}_a^{BLUP} = \varphi_a\left(\bar{y}_a + (\bar{\mathbf{X}}_{\boldsymbol{a}} - \bar{\mathbf{x}}_{\boldsymbol{a}})^{\mathrm{t}}\hat{\boldsymbol{\beta}}\right) + (1 - \varphi_a)(\bar{\mathbf{X}}_{\boldsymbol{a}}^{\mathbf{t}}\hat{\boldsymbol{\beta}}), \tag{2}$$

where $\hat{\boldsymbol{\beta}}$ is the vector of fixed effects estimated at the national level and φ_a is a weight between the direct and synthetic estimator given by $\varphi_a = \sigma^2/(\Psi_a^2 + \sigma^2)$. Now, if in Equation (2), the variance of the random area effects σ^2 is replaced by its estimator $\hat{\sigma}^2$, the empirical BLUP (EBLUP) estimator is obtained. Moreover, the sampling variance Ψ_a^2 is assumed to be known; however, in practice, this is not true and, in this application, it is replaced by its estimator obtained with the paired jackknife. The mean squared error

(MSE) of the EBLUP accounts for the additional uncertainty that is introduced, since σ^2 is replaced by its estimator $\hat{\sigma}^2$ but ignores the uncertainty of using an estimator for Ψ_a^2, which is common practice in SAE procedures.

In this article, an HB approach is applied to fit Equation (2). The HB model is based on Equation (1) under the assumption that $e_a \sim N(0, \psi_a^2)$ and $u_a \sim N(0, \sigma^2)$. For $\boldsymbol{\beta}$ and σ^2, a flat prior distribution is assumed. The HB estimates for the area means, including their MSEs, are obtained by the posterior means and posterior variances of the posterior density for the area means μ_a. These estimates can be evaluated using separate one-dimensional numerical integrations.

To obtain stable variances for the survey regression estimates, the variance approximations obtained with the jackknife are pooled using an analysis of variance type pooled estimator:

$$\Psi_a^{2;P} = \frac{1}{N_a} \frac{\Sigma_{a=1}^{m}(N_a - 1)\Psi_a^2}{\Sigma_{a=1}^{m}(N_a - 1)},$$

where m is equal to the total number of areas.

Furthermore, it was clear that some municipalities had unrealistically low literates estimates (one was even negative): they were underestimated due to the linearity of the model. Therefore, two post-result changes were implemented. First, we acknowledged that the model had problems estimating the true percentages in areas where the percentage of low literates is very small ($<5\%$), which is further considered in the results. So, during categorization, we marked these municipalities as having a very small percentage (0–5%) of low literates and grouped them together when publishing the results. Second, a choice was made to benchmark the results such that they would add up to the national level as per You et al. (2004), by means of the direct estimate of undercoverage per area and the sampling variances.

Since the dependent variable in the Fay–Herriot model are direct estimates of percentages, we also considered a log odd transformation, that is, Equation (1) applied to $\log\left(\hat{y}_a^{surv}/(1-\hat{y}_a^{surv})\right)$. As shown in the results section, the area level model after applying a log-odds transformation results in more biased domain predictions than the area level model applied to the untransformed estimates. 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. 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).

3.3. *Unit-Level Model*

As before, let y_{ia} denote the average of the 10 plausible values of the literacy proficiency level of an individual i in area a. The true mean is then equal to

$$y_{ia} = \mu_{ia} + e_{ia} = \alpha + \mathbf{x}_{ia}^{t}\boldsymbol{\beta} + u_a + e_{ia}, \tag{3}$$

where $\mathbf{x}_{ia}$ is a vector with covariates for respondent i from area a and u_a is an area-specific random effect assumed to be independent and identically distributed. We assume e_{ia} is a

measurement error for respondent i, with expected value zero and variance σ_e^2. The EBLUP estimator is then equal to

$$\hat{y}_a^{EBLUP} = \varphi_a(\hat{y}_a^{surv}) + (1 - \varphi_a)(\bar{\mathbf{X}}_a^{\mathbf{t}}\hat{\boldsymbol{\beta}}),$$

where the weight φ_a, dependent on area size N_a, is given by $\varphi_a = \sigma^2/(\sigma^2 + \sigma_e^2/N_a)$. The HB model is obtained with Equation (3) with the assumption that $e_{ia} \sim N(0, \sigma_e^2)$ and $u_a \sim N(0, \sigma^2)$. Furthermore, flat priors are assumed for $\boldsymbol{\beta}$, σ_e^2, and σ^2. The HB predictors for the area means, for example, $\hat{y}_a^{HB}$, with their MSEs, are computed as the posterior means and posterior variance of the posterior distribution of μ_a in a similar way as for the area-level model. The resulting integrals are solved using numerical integration.

Unlike the area-level model for the percentage of low literates, where the imputation uncertainty is taken into account when constructing $\bar{y}_a$, the unit-level model as described above does not take into account the imputation uncertainty.

Multiple imputation is one way to take into account this imputation uncertainty, combining results by means of Rubin's rules (Rubin, 1996). The plausible values generated with the PIAAC software are used to calculate multiple HB predictions for the areas. Let $\hat{y}_{aj}^{HB}$ denote the HB prediction for area a based on the jth set of plausible values generated for the PIAAC sample and $MSE(\hat{y}_{aj}^{HB})$ denote the posterior variance of $\hat{y}_{aj}^{HB}$. The final HB prediction for area a is defined as

$$\hat{y}_a^{imp} = \sum_{j=1}^{k} \frac{\hat{y}_{aj}^{HB}}{k},$$

where k is the total number of plausible values. The total variance V_a^{imp} is equal to

$$V_a^{imp} = W_a + \frac{k+1}{k} B_a,$$

where the within-imputation variability W_a is obtained as the mean over the MSE of the HB small area predictions:

$$W_a = \sum_{j=1}^{k} \frac{MSE(\hat{y}_{aj}^{HB})}{k}.$$

The between-imputation variability B_a is

$$B_a = \sum_{j=1}^{k} \frac{(\hat{y}_{aj}^{HB} - \hat{y}_a^{imp})^2}{k-1}.$$

Note that Rubin's rule for multiple imputation is derived for large samples. It is unclear to what extent the application of this methodology to small area estimation problems introduces additional bias in point estimates and uncertainty measures. This is left for further research.

4. Model Fitting

4.1. Merging of Municipalities

As stated before, in 2012 the Netherlands was comprised of 415 municipalities. However, some municipalities are quite small and we cannot guarantee that their LFS data cover

enough respondents to provide an accurate representation of its inhabitants. Therefore, it is necessary to work with municipality clusters instead. We use 40,000 as the minimum number of residents per area to ensure the LFS estimates can be considered reliable, for example, the variance being low enough to be negligible. This minimum value is based on Statistics Netherlands' publication strategy that three year averages of direct LFS estimates are published for municipalities with a minimum of 40,000 residents aged 16 years and over from 2010 onwards. Municipalities with fewer residents are clustered together with adjacent municipalities. During this merging, we made sure that all the areas could still be nested in larger official area aggregates, the COROP regions. This is a 40-area classification based on educational provisions. Finally, 208 municipality clusters are obtained, for which small area estimates about literacy will be made. In the PIAAC sample, the minimum number of observations for these clusters is 6, the maximum is 146, and the median is 20.

4.2. *Variable Selection*

SAE uses auxiliary variables at the area level for additional predictive power. This means that all data available in the LFS that is also included in the PIAAC questionnaire can be picked for use in our model. The list of auxiliary variables for the full model and descriptive results (averages and standard deviations) are presented in Table 2.

Table 2. Comparison of weighted dataset averages and their standard deviations (in parentheses).

Covariate[1]	PIAAC average[2]	LFS average
Age**[4]	41.0 (14.2)	40.6 (14.1)
Male	49.3% (50.0)	50.2% (50.0)
ISEI08-score***	48.7 (18.4)	46.5 (10.6)
Immigrant status		
1st gen***	12.8% (32.6)	14.0% (34.7)
2nd gen***	3.1% (16.8)	9.4% (29.2)
Employment status		
Student	13.9% (34.4)	13.7% (33.8)
Self-employed	9.1% (28.7)	9.1% (29.8)
Full time employee***	37.5% (48.4)	30.9% (46.2)
Part time employee	22.1% (41.5)	21.6% (41.2)
Unemployed***	2.6% (16.0)	3.5% (18.4)
Education[3]		
Vocational ed.	57.5% (49.4)	57.5% (49.4)
Years of schooling***	13.2 (3.7)	13.4 (3.6)

[1]The full list of interactions considered for the full model are age with gender, ISEI-08 score, immigrant status variables, employment status variables and education variables, plus years of schooling with immigrant status variables, ISEI-08 score and vocational education.

[2]For the Netherlands, the control variables that were used to calibrate weights in PIAAC are: Gender by age (10), origin by generation (5), group of provinces by degree of urbanization (18), household type (5), social status by income (25), term of registration in population registry (2), percentage of high level education by percentage of low level education (18).

[3]The education variables contained slightly more than 1% missing values. For area estimates, missing values are assumed have the same distribution as the known values.

[4]Indicates the level of statistical significance of the t-test between the two datasets. ***$p < 0.001$, **$p < 0.05$, *$p < 0.01$.

There are some statistically significant differences in the distribution of these variables between PIAAC and LFS, although most of these differences in distribution are rather small in nature; our large sample sizes allow even minor differences to be statistically significant. The most notable difference is the percentage of second-generation immigrants in the PIAAC data set, which is significantly lower in the PIAAC data set compared to the LFS data set. Also, there is a (non-significant) larger percentage of fulltime employees, and a lower percentage of unemployed persons. There are some minor differences for age, occupational status and years of schooling where the gap between the means is very small.

In the literature, different methods are proposed for model selection. In this article, optimal models are selected by means of the conditional Akaike information criterion (cAIC) using a stepwise backward variable selection procedure. This method is applied more often in small area estimation (see e.g., Van den Brakel and Buelens 2015). The cAIC, proposed by Vaida and Blanchard (2005), is applicable to mixed models where the focus is on prediction at the level of areas. The penalty (p) on the log likelihood is based on the model complexity. The random part of the model contributes to the number of degrees of freedom p with a value between zero in the case of no area effects (i.e., $\hat{\sigma}^2 = 0$) and the total number of areas m in the case of fixed area effects (i.e., $\hat{\sigma}^2 \rightarrow \infty$). The effective number of degrees of freedom used for the penalty is defined as the trace of the hat matrix H, which maps the observed data to the fitted values, for example $\hat{y} = Hy$, see Hodges and Sargent (2001). The cAIC has a more realistic penalty for the random component of a multilevel model, compared to the standard AIC (where a random effect counts for one degree of freedom). Nevertheless, the cAIC in a stepwise selection procedure might result in complex models that overfit the data. Alternatively, cross-validation is sometimes used as a measure for model selection, see Boonstra et al. (2008). Other authors propose the LASSO (Hastie et al. 2001) as a form of model selection (Thao and Geskus 2019). In this article, the cAIC is used in combination with a backward selection procedure and in the model evaluation it is established that the selected models do not overfit the data.

Covariates were removed one by one until a minimum for the cAIC was reached for the unit-level model on literacy scores. The list of the selected predictors is as follows:

- Age, Age squared,
- Immigrant Status,
- Years of Schooling,
- Area of Study (eight categories),
- Highest level of education is Vocational Education (Dummy); Note that vocational education in the Netherlands can be secondary, upper-secondary and tertiary level,
- Employment Status,
- Occupational Status Measure based on the International Socio-Economic Index (ISEI) of ISCO-08 occupations by Ganzeboom et al. (1992), a continuous variable measuring the socio-economic status of an occupation,
- Two 2-way interaction terms of Years of Schooling with Immigrant Status and Occupational Status, and
- Six 2-way interaction terms of Age with Gender, Vocational Education and Employment Status.

The interaction terms help with estimating effects of variables not captured in our data sets. For example, international knowledge workers would be classified as immigrants, which is generally a negative indicator. By including the interaction effect with years of schooling, we can partially correct for this. For the area level model, we can find a model with a slightly lower cAIC score (ΔcAIC $= 2.9$) by leaving out the self-employed and one dummy regarding the area of study. However, in theory there is no reason why the two sets of literacy measures should have different predictors. Given the small difference in model selection, we opt to use the same model for both predictors. A quick test using the other model reveals that all results lie within the confidence interval of our preferred model.

5. Model Evaluation

The SAE results can differ from the direct results for a number of reasons. The most important reason is why SAE techniques are applied in the first place, namely, to improve the precision of the direct municipality estimates. However, it is important to make sure the differences are not dominated by the bias introduced in the model. Since SAE techniques explicitly rely on statistical models to improve the effective sample size in the separate areas, one must evaluate the underlying assumptions of the models to ensure the bias introduced by the synthetic estimator is small. Model misspecification can easily result in heavily biased area estimates. This section evaluates the normality assumptions underlying the applied models. Furthermore, direct area estimates are compared with model-based small area predictions to assess possible systematic bias. Finally, the improvement in precision is evaluated by comparing the standard errors of both estimators.

5.1. Robustness Checks

The direct estimates at the national level are precise and unbiased, since they do not depend on model assumptions and are based on a large sample. Therefore, the difference between the model-based small area predictions, aggregated at the national level, with the direct estimates at the national level is often used as a measure of bias in SAE.

As noted earlier in Section 3, benchmarking was applied to remove differences between model-based area estimates aggregated at the national level and direct estimates at the national level. Small area estimates for literacy scores and the percentage of low literates at the national level are obtained by calculating the mean over the municipalities weighted by the number of residents in 2012. Table 3 displays the results of the non-benchmarked estimates against the (robust) national results.

Table 3. Estimated aggregated results at higher levels, without benchmarking.

Type	Direct	SAE (*)
Average Literacy	283.9	287.9
% Low Literates	12.0%	12.8%

*indicates the average of the SAE results over municipalities, weighted by the number of residents in 2012.

For both measures of literacy, the SAE scores are slightly overestimated. The average literacy of 287.9 is greater than the upper bound of 285.3 for the direct estimates given in Table 1. The estimate of the percentage of low literates estimates is contained within the 95% confidence interval, but barely. On the basis of these results, we decided to benchmark our estimates.

Before benchmarking, we look at the differences between the direct estimates and the SAE results. Two measures are applied to summarize the differences between the direct and model-based area estimates. The first one is the *mean relative difference* (MRD), in percentages, defined as

$$MRD = \frac{1}{m}\sum_{a=1}^{m} \frac{(\hat{y}_a^{direct} - \hat{y}_a^{SAE})}{\hat{y}_a^{direct}} 100,$$

where $\hat{y}_a^{SAE}$ is the unbenchmarked Hierarchical Bayesian SAE estimator. The second one is the *absolute mean relative difference* (AMRD), in percentages, defined as

$$AMRD = \frac{1}{m}\sum_{a=1}^{m} \frac{(\hat{y}_a^{direct} - \hat{y}_a^{SAE})}{\hat{y}_a^{direct}} 100.$$

Table 4 gives the MRD and AMRD for the two literacy measures.

The MRD for both estimates is quite small, with roughly 1.7 percentage point for the average literacy and half a percentage point for the low literacy percentage. Since it is negative, the SAE estimators are generally slightly bigger. When we look at the absolute difference, we see a 2.78% mean difference for average literacy, and 0.70% for low literacy.

To interpret the differences between the direct estimates and the domain predictions obtained with the finally selected SAE models in more detail, we compare the distribution of the benchmarked SAE estimates with the distribution of the direct results from PIAAC. Figure 1 shows the tendency of the SAE estimates to tend towards the mean. Regarding the average literacy scores, the scores at the right side of the distribution consist mostly of those for university cities, where the number of students seems to be oversampled. The scores at the left side of the distribution are mostly for small villages, but the worst results are for some municipalities of medium-sized cities.

For the estimated percentage of low literates, the distribution is close to the distribution of the direct estimates; however, note that the SAE results for the average and below-average percentage of low literates are often higher than the direct results. The relatively high proportion of municipalities (over 10%) that perform well in terms of percentage of low literates (with percentages in the range of 0–5%) in the direct estimates could be due to the fact that these municipalities are very small and have few direct observations in

Table 4. Measures of quality of the estimates (%), without benchmarking.

	Average Literacy	% Low Literates
MRD	−1.66	−0.51
AMRD	2.78	0.70

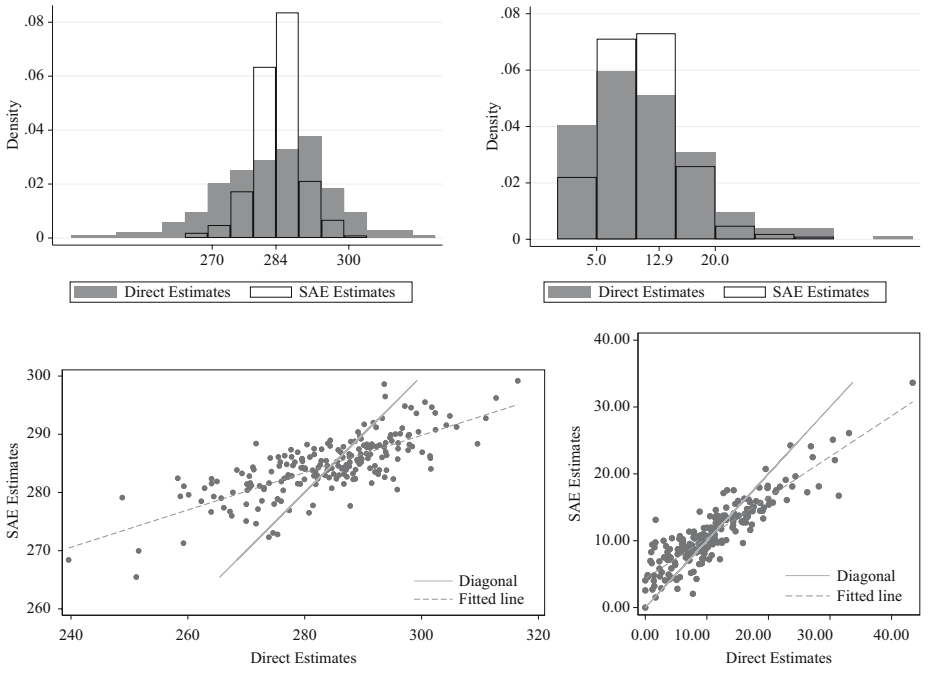


Fig. 1. Histograms and distribution plots of the direct results and the SAE results (left, literacy scores; right, % low literates; the solid line is the diagonal, the dashed line is the linear fit).

PIAAC. Therefore, these differences would be a result of the improved accuracy of the point estimates.

Figure 2 shows the scatter plots of the fitted values of both SAE measures versus the quantiles of the residuals. No pattern can be distinguished within the two graphs, meaning the residuals are well behaved.

Q-Q plots for the estimates, residuals and random effects can be found in the Supplementary materials.

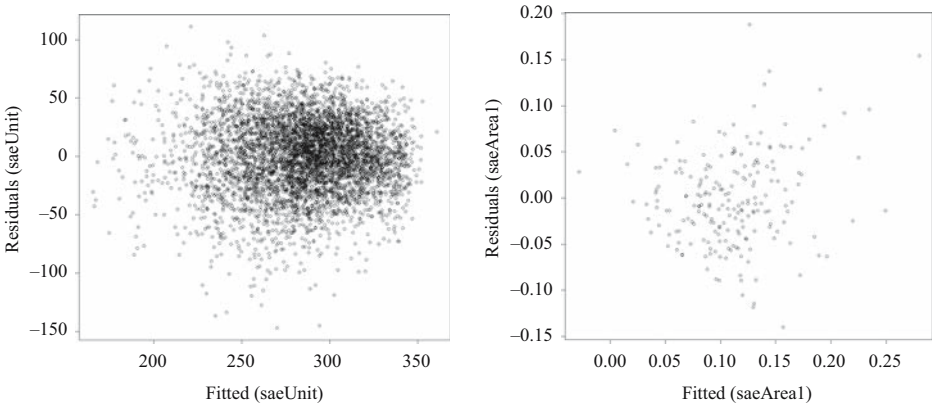


Fig. 2. Fitted values versus the residuals of the unit-level estimates of the estimated literacy scores (left) and the area-level estimates of the percentage of low literates after benchmarking (right).

For the percentage of low literates, a log odds transformation of the dependent variable was also considered and applied. The model under the log odds transformation shrinks in particular the direct domain estimates with small values much stronger to the overall mean, resulting in larger amounts of bias (RMD and ARMD have values of respectively -1.485 and 1.580). Furthermore, the residuals and random effects show stronger deviations from normality. See the Supplementary materials for more details. Therefore, the model applied to the untransformed direct estimates is chosen to be our final model. As explained in Section 3, this is not uncommon in survey sampling and SAE literature.

5.2. *Reduction in Standard Error*

To measure the increase in precision obtained with the SAE techniques, the *mean relative difference in standard errors* (MRDSE) is used. This is defined as the ratio between the standard errors between the direct and the SAE estimator, averaged per area, or in formula form:

$$MRDSE = \frac{1}{m}\sum_{a=1}^{m} \frac{(SE_a^{direct} - SE_a^{Bench})}{SE_a^{direct}} * 100$$

The results are shown in Table 5. The MRDSE for average literacy is 67.9%, which, compared to the direct estimates, is a significant reduction. For the percentage of low literates, the reduction measure is 51.2% (31.3%) when compared to the pooled variance) but, as a less powerful model, lower returns are to be expected.

In Figure 3, we look at the number of respondents in PIAAC versus the standard error of the direct estimates, as well as the SAE results for the average literacy scores per municipality. Given the high frequency of respondents numbering between 5 and 20 per municipality, we decided to plot this graph on a logarithmic scale.

For small sample sizes, the SAE results show a large decrease in terms of standard errors compared to the direct estimator, whose margin of error is far too large when it comes to accurate point estimates. As the sample size increases, the difference between the two estimators decreases greatly.

In Figure 4, we look at the standard errors for the percentage of low literates. Here, the standard errors of the direct estimator are much more spread out and sometimes even zero (due to the direct estimator being zero). When compared to the direct estimator with pooled standard errors they are much closer to the SAE results due to the decrease in information compared to the model utilizing literacy scores, but there is still a significant gain in municipalities with low numbers of PIAAC respondents.

Table 5. Measures of the quality of estimates (%), without benchmarking.

	Average Literacy	% Low Literates*
MRDSE	67.9	51.2 (31.3)

*indicates the numbers in parentheses are compared to the standard errors of the pooled variance instead of the direct standard errors.

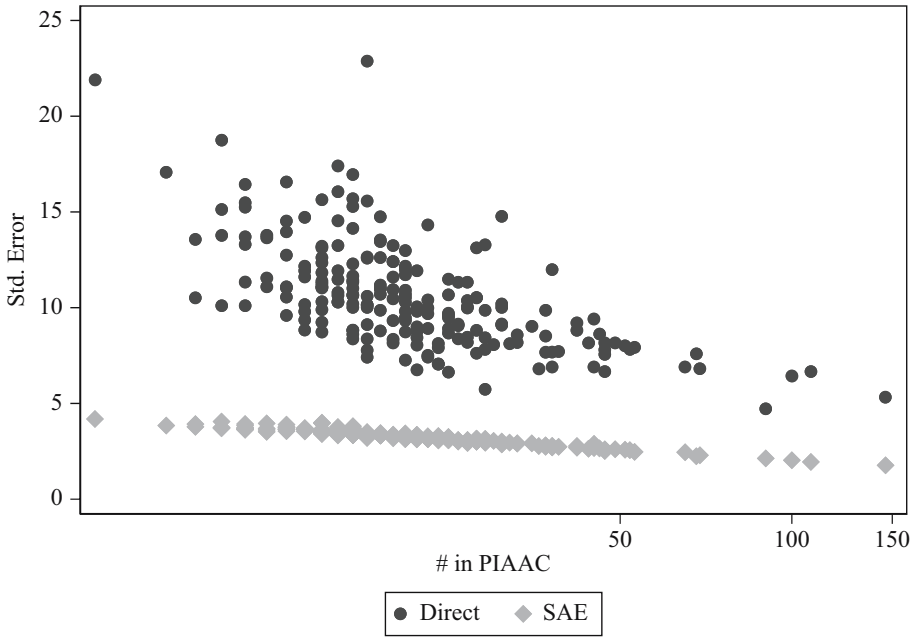


Fig. 3. Standard errors versus the (logarithmic) number of PIAAC respondents for both the direct estimates and the SAE for the estimated literacy scores per municipality.

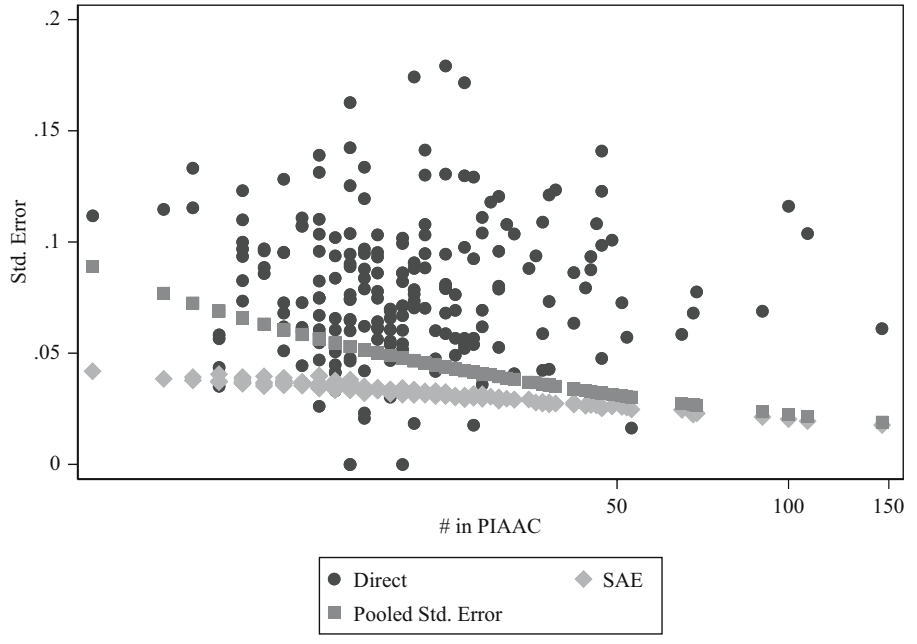


Fig. 4. Standard errors versus the (logarithmic) number of in PIAAC respondents for both estimates and the SAE for the percentage of low literates per municipality.

6. Results

In this section, we present the substantive results graphically, review them, and discuss the differences in results for the two chosen measures of literacy. The full list of results per municipality can be found in the online Supplementary material.

Figure 5 shows the average literacy scores per municipality cluster. Neighbors are rarely in the same category and often differ by multiple categories. Generally, the highest scores for literacy can be found in the center of the country, around the city of Utrecht. Large university cities also do well (Rotterdam being a notable exception). Aside from known problem areas in the western part of the Netherlands, the scores for literacy are low in the peripheral regions.

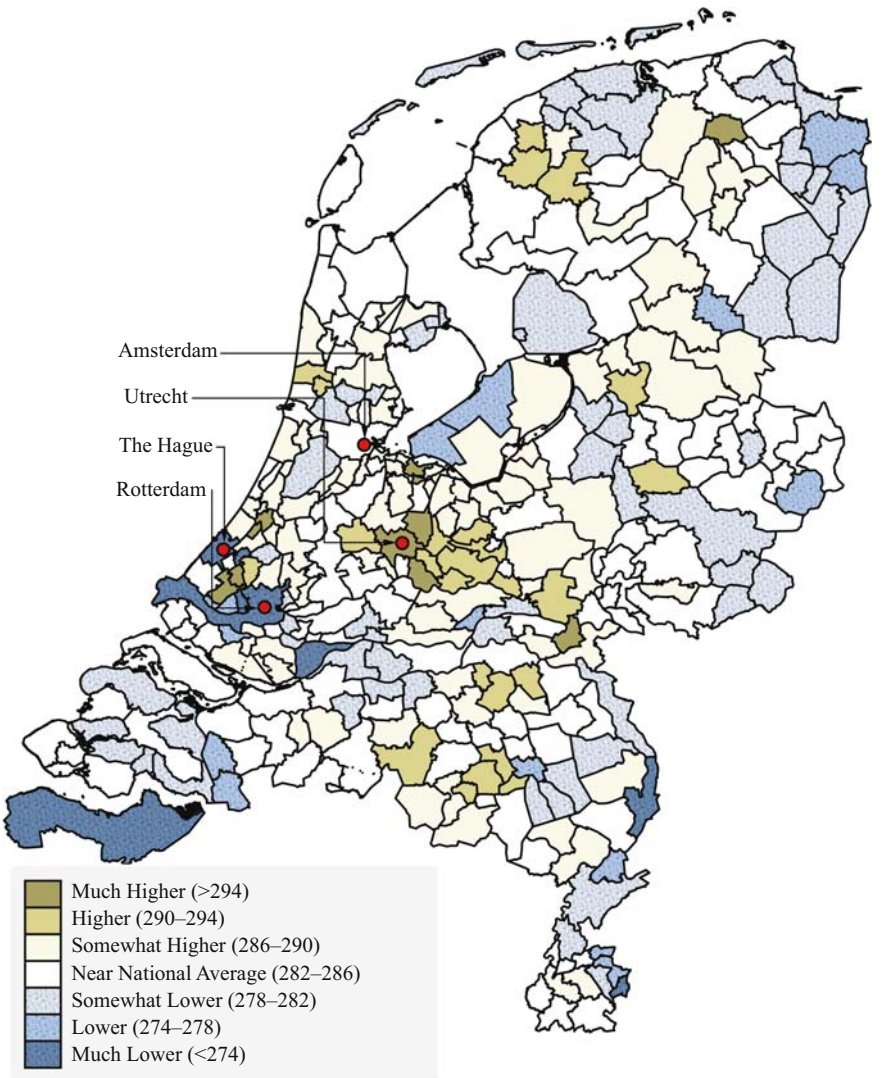


Fig. 5. Estimated average literacy scores per municipality.

Figure 6 shows the regional estimates for the percentage of low literates. There is a similar pattern when we look at areas in terms of the percentage of low literates. The first big notable difference, however, is that, in most cases, large cities do much worse in terms of their percentage of low literates in their population, which underlines the usefulness of having both indicators. Low literacy is mainly found in populations with certain characteristics. The average literacy score could give an idea of the overall situation of a population, but not how it is distributed. Both measures together provide a more complete picture of the literacy within each area.

Next, we give some examples of how SAE estimates for literacy can relate to other outcomes at the regional level. Knowledge of regional differences can be a powerful tool

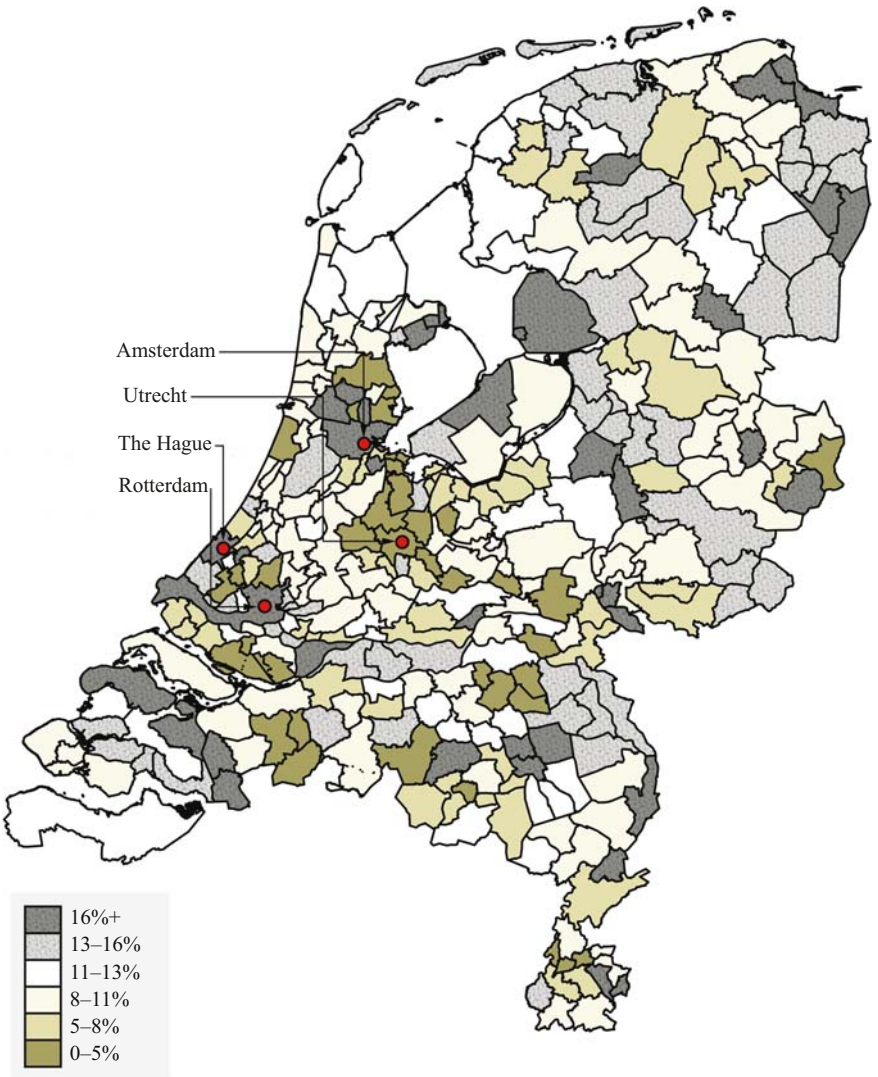


Fig. 6. Estimated percentage of individuals classified as having low literacy proficiency scores per municipality.

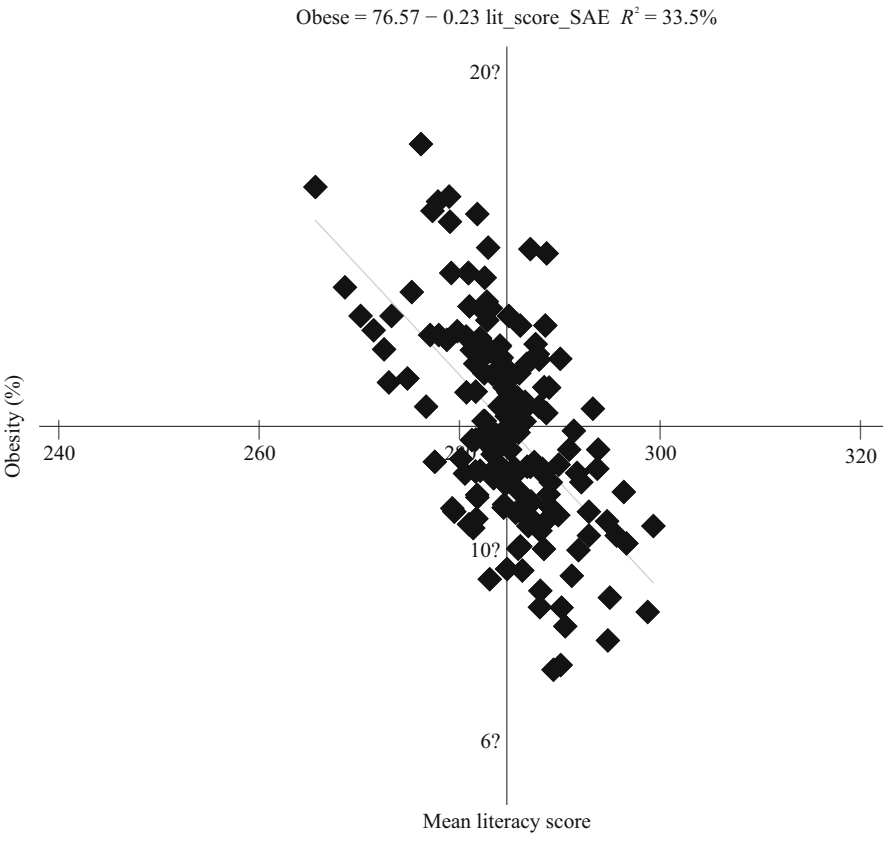


Fig. 7. Linear model of the proportion of obese people (in 2012; Source: Statistics Netherlands) versus the average literacy estimates in that region.

for policy interventions aimed at tackling these problems. This is not simply a matter of identifying areas of low literacy, since this is unlikely to be the sole cause of such problems. Policy makers and professionals responsible for policy implementation have an interest in distinguishing regions in which poor health, and other unwanted outcomes are associated with low literacy from regions in which these problems are driven more by other factors. Such knowledge can greatly improve the cost effectiveness of interventions.

As a simple illustration, in Figure 7, we plot the relation between (low) literacy and one unwanted non-economic problem: obesity. Note that the following is for illustration purposes only. This approach facilitates the implementation of more targeted policy interventions. The idea behind this is the following. Very often problems like low literacy, health problems or socio-economic problems go hand in hand. Policy, therefore, is often aimed at an integral approach, such as a combination of helping to find work, improvement of a healthy lifestyle and improving the literacy proficiency. For policy makers it is helpful to see which combinations of problems occur in their municipality so that they can fine-tune their interventions for the specific group. Our goal is not to 'explain' obesity, but to

identify areas in which there is an accumulation of both types of problems versus areas where this is not the case.

The relation between the average literacy score and the incidence of obesity is quite strong ($R^2 = 33.5\%$), but also far from perfect. There are areas where the two problems go hand in hand and areas where this is not the case at all. In terms of policy interventions, the position of a given municipality in the graph is indicative of the kind of policy response that could be considered appropriate. There is little incentive to launch literacy-based interventions in the regions in the lower right quadrant, since these are regions with high literacy and a low incidence of obesity. In the lower left and upper right quadrants, literacy-based interventions also do not look promising, at least not to combat obesity, since literacy and obesity do not coincide in these regions. Only in the upper left quadrant do we see a high incidence of obesity together with a low average level of literacy. This finding suggests that literacy could potentially be targeted as a policy lever to tackle the problem of obesity in these regions.

7. Conclusion

In this article, we have combined PIAAC survey data with LFS data to obtain estimates of the literacy levels for municipalities in the Netherlands, both the average literacy scores and the percentage of low literates. These estimations are obtained using SAE models fitted with an HB approach.

Direct estimators only use observations obtained in each specific area to estimate literacy for that area. Results obtained with direct estimators at the regional level, therefore, suffer from small samples sizes for most areas, leading to high standard errors. In this article, we applied model-based estimation procedures to improve the effective sample size in the different areas, resulting in a considerable improvement of the precision of the estimates of literacy levels, even in larger cities of the Netherlands.

We show that we can obtain estimates at a very detailed regional level by using these SAE techniques, with standard errors reduced more than 50%. This is important, since policy to combat low literacy is often targeted at the municipality level. We show that we can obtain reliable estimates for the average literacy level and the percentage of low literates for over 200 municipalities in the Netherlands. The findings show that average literacy levels are higher in big cities than in more rural areas, a finding that is consistent with the literature (e.g., McHenry 2014). However, we also show that large cities cope with higher proportions of low literates, indicating the importance of looking at both measures of literacy.

The estimates can help to determine a more optimal allocation of resources to combat low literacy. We also illustrated that more precise SAE estimates are helpful in establishing relations with other variables more clearly. This approach can be used, for example, to identify municipalities that suffer from multiple problems, such as low literacy and health problems or other social problems. In some municipalities, these problems coincide, and in some municipalities they do not. Identifying the typical mix of problems a municipality is confronted with is key to the development of a successful intervention strategy. The regional estimates for literacy, therefore, give room for policy makers to implement more directed policies at a detailed regional level.

Future research will focus on the estimation of other skills measured in PIAAC, such as numeracy, or by estimating literacy levels in other areas, such as detailed levels of occupation (for an example, see Van der Velden and Bijlsma 2018). By making these kinds of estimates possible, detailed data become available in areas previously inaccessible due to time and budget constraints.

However, there are a number of caveats to keep in mind when interpreting the results. First and foremost, it must be stressed that these methods rely on statistical model assumptions. Careful model selection and evaluation are, therefore, an important and necessary part of SAE. The method assumes that the effects of covariates at the regional level are the same as at the national level, with random effects capturing regional differences. While this should hold in most cases, exceptions can occur. The results should always be viewed with possible local anomalies in mind.

A number of improvements can be made in the estimation of the model. Currently, data used from the LFS are assumed to be the true population means and the corresponding sampling errors are assumed to be negligible. There are ways to properly consider these errors, such as the method of Ybarra and Lohr (2008) for the area-level model and the method of Lohr and Prasad (2003) for the unit-level model. For the percentage of low literates model, a logarithmic model could lead to better estimations between the 0% and 5%, which currently show some bias toward the bottom end of the distribution. Methods such as the standard ratio raking used in Casas-Cordero et al. (2016) are also an option.

8. 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://doi.org/10.1111/rssa.12321.

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* 401: 28–36. DOI: https://doi.org/10.1080/01621459.1988.10478561.

Boonstra, H.J. 2015. *Package 'hbsae'* (version 1.0). Available at: https://cran.r-project.org/web/packages/hbsae/hbsae.pdf (accessed December 2015).

Boonstra, H.J., J.A. van den Brakel, B. Buelens, S. Krieg, and M. Smeets. 2008. "Towards small area estimation at Statistics Netherlands." *METRON International Journal of Statistics* LXVI: 21–49. Available at: https://EconPapers.repec.org/RePEc:mtn:ancoec:080102 (accessed April 2020).

Buisman, M., J. Allen, D. Fouarge, W. Houtkoop, and R. van der Velden. 2013. *PIAAC: Kernvaardigheden voor werk en leven. Resultaten van de Nederlandse survey 2012*, Den Bosch/Maastricht: ECBO/ROA.

Casas-Cordero, C., J. Encina, and P. Lahiri. 2016. "Poverty mapping for the Chilean Comunas." In *Analysis of Poverty Data by Small Area Estimation*, edited by M. Pratesi, 379–403. Hoboken: Wiley. DOI: https://doi.org/10.1111/j.1467-9787.2007.00538.x

Coulombe, S. and J.F. Tremblay. 2007. "Skills, Education, and Canadian Provincial Disparity." *Journal of Regional Science* 47: 965–991. DOI: https://doi.org/10.2307/2669921.

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* 448: 1074–1082.

Elbers, C., J.O. Lanjouw, and P. Lanjouw. 2003. "Micro estimation of poverty and inequality." *Econometrica* 71: 355–364. DOI: https://doi.org/10.1111/1468-0262.00399.

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* 366: 269–277. DOI: https://doi.org/10.2307/2286322.

Ganzeboom, H.B.G., P.M. de Graaf, and D.J. Treiman. 1992. "A Standard International Socio-Economic Index of Occupational Status." *Social Science Research* 21: 1–56. DOI: https://doi.org/10.1016/0049-089X(92)90017-B.

Gibson, A. and P. Hewson. 2012. "2011 Skills for Life Survey: Small Area Estimation Technical Report." *BIS Research Report* 81C. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/36077/12-1318-2011-skills-for-life-small-area-estimation-technical.pdf (accessed November 2018).

Hanushek, E.A. and L. Woessmann. 2008. "The Role of Cognitive Skills in Economic Development." *Journal of Economic Literature* 46: 607–668. DOI: https://doi.org/10.3386/w15949.

Hanushek, E.A. and L. Woessmann. 2011. The Economics of International Differences in Educational Achievement. In *Handbook of the Economics of Education*, Vol. 3: 89–200. Amsterdam: North Holland.

Hastie, T., R. Tibshirani, and J. Friedman. 2001. *The elements of statistical learning*. Springer: New York.

Hodges, J.S. and D.J. Sargent. 2001. "Counting degrees of freedom in hierarchical and other richly parameterized models." *Biometrika* 88: 367–379. DOI: https://doi.org/10.1093/biomet/88.2.367.

Johnson, E.G. and K.F. Rust. 1992. "Sampling and Weighting in the National Assessment." *Journal of Educational and Behavioral Statistics* 17: 111–129. DOI: https:/doi.org/10.2307/1165165.

Lohr, S. and N. Prasad. 2003. "Small Area Estimation with Auxiliary Survey Data." *The Canadian Journal of Statistics* 31: 383–396. DOI: https://doi.org/10.2307/3315852.

McHenry, P. 2014. "The Geographic Distribution of Human Capital: Measurement of Contributing Mechanisms." *Journal of Regional Science* 54: 215–248. DOI: https://doi.org/10.1111/jors.12067.

National Research Council. 2000. "Small Area Estimates of School-Age Children in Poverty: Evaluation of current methodology." *Committee on National Statisitcs*, edited by C.F. Citro and G. Kalton. Washington, DC: National Academy Press.

OECD. 2013a. *OECD skills outlook 2013: first results from the survey of adult skills*. Paris: OECD Publishing. DOI: https:/doi.org/10.1787/9789264204256-en.

OECD. 2013b. *The Survey of Adult Skills – Reader's Companion*. Paris: OECD Publishing. DOI: https://doi.org/10.1787/9789264204027-en.

OECD. 2013c. *Technical Report of the Survey of Adult Skills (PIAAC)*. Available at: http://www.oecd.org/site/piaac/publications.htm (accessed December 2015).

Pokropek, A. and M. Jakubowski. 2013. *Package 'PIAAC tools'* (version 4.3). Available at: https://ideas.repec.org/c/boc/bocode/s457728.html (accessed September 2016).

Pfeffermann, D. 2013. "New Important Developments in Small Area Estimation." *Statistical Science* 28: 40–68. DOI: https://doi.org/10.1214/12-STS395.

PricewaterhouseCoopers. 2013. *Laaggeletterdheid in Nederland kent aanzienlijke maatschappelijke kosten.* Internal Rapport, PWC, Amsterdam.

Rao, J.N.K. and I. Molina. 2015. *Small Area Estimation, Second Edition*. New York: John Wiley and Sons.

Rubin, D.B. 1996. "Multiple Imputation After 18 + Years." *Journal of the American Statistical Association* 434: 473–489. DOI: https://doi.org/10.2307/2291635.

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

Schmid, T., F. Bruckschen, N. Salvati, and T. Zbiranski. 2017. "Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in Senegal." *Journal of the Royal Statistical Society Series A (Statistics in Society)* 180: 1163–1190. DOI: https://doi.org/10.1111/rssa.12305Y.

Statistics Netherlands. 2010. "Methoden en definities Enquête Beroepsbevolking 2010." Available at: https://www.cbs.nl/nl-nl/onze-diensten/methoden/onderzoeksomschrijvingen/aanvullende%20onderzoeksbeschrijvingen/enquete-beroepsbevolking-uitgebreide-onderzoeksbeschrijving-2010 (accessed March 2018).

Statistics Netherlands. 2011. "Methoden en definities Enquête Beroepsbevolking 2011." Available at: https://www.cbs.nl/nl-nl/onze-diensten/methoden/onderzoeksomschrijvingen/aanvullende%20onderzoeksbeschrijvingen/enquete-beroepsbevolking-uitgebreide-onderzoeksbeschrijving-2011 (accessed March 2018).

Statistics Netherlands. 2012. "Methoden en definities Enquête Beroepsbevolking 2012." Available at: https://www.cbs.nl/nl-nl/onze-diensten/methoden/onderzoeksomschrijvingen/aanvullende%20onderzoeksbeschrijvingen/enquete-beroepsbevolking-uitgebreide-onderzoeksbeschrijving-2012 (accessed March 2018).

Taylor, J., G. Moon, and L. Twigg. 2016. "Using geocoded survey data to improve the accuracy of multilevel small area synthetic." *Social Science Research* 56: 108–116. DOI: https://doi.org/10.1016/j.ssresearch.2015.12.006.

Thao, L.T.P. and R. Geskus. 2019. "A comparison of model selection methods for prediction in the presence of multiply imputed data." *Biometrical Journal* 61: 343–356. DOI: https://doi.org/10.1002/bimj.201700232.

Tighe, E., D. Livert, M. Barnett, and L. Saxe. 2010. "Cross-Survey Analysis to estimate low-incidence religious groups." *Sociological Methods & Research* 39: 56–82. DOI: https://doi.org/10.1177/0049124110366237.

Vaida, F. and S. Blanchard. 2005. "Conditional Akaike information for mixed effect models." *Biometrika* 92: 351–370. DOI: https://doi.org/10.1093/biomet/92.2.351.

Van den Brakel, J.A. and B. Buelens. 2015. "Covariate selection for small area estimation in repeated sample surveys." *Survey Methodology and Statistics in Transition*, Special issue on *Small Area Estimation,* Vol.16: 523–540. DOI: https://doi.org/10.21307/stattrans-2015-031.

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: https://www150.statcan.gc.ca/n1/pub/12-001-x/2015002/article/14231-eng.pdf (accessed April 2020).

Van der Velden, R. and I. Bijlsma. 2018. "Effective skill: a new theoretical perspective on the relation between skills, skill use, mismatches and wages." *Oxford Economic Papers*, Advance articles. DOI: https://doi.org/10.1093/oep/gpy028.

World Bank. 2002. "How Low Can You Go? Combining Census and Survey Data for Mapping Poverty in South Africa." *Journal of African Economies* 11: 169–200. DOI: https://doi.org/10.1093/jae/11.2.169.

Yamamoto, K. 2014. *Using PIAAC Data for Producing Regional Estimates*. Working Paper, Educational Testing Service, Princeton.

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.N.K. Rao, and P. Dick. 2004. "Benchmarking Hierarchical Bayes Small Area Estimators in the Canadian Census Undercoverage Estimation." *Statistics in Transition* 6: 631–640. Available at: https://www.semanticscholar.org/paper/BENCHMARKING-HIERARCHICAL-BAYES-SMALL-AREA-IN-THE-You-Rao/efaafa565aa134fe0943f03bbad15278eb228e3a (accessed April 2020).

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/en/catalogue/12-001-X20030016602 (accessed April 2020).

Received November 2018
Revised August 2019
Accepted January 2020

sciendo

Journal of Official Statistics, Vol. 36, No. 2, 2020, pp. 275–296, http://dx.doi.org/10.2478/JOS-2020-0015

Analysing Sensitive Data from Dynamically-Generated Overlapping Contingency Tables

Joshua J. Bon[1], *Bernard Baffour*[2], *Melanie Spallek*[3], *and Michele Haynes*[3]

Contingency tables provide a convenient format to publish summary data from confidential survey and administrative records that capture a wide range of social and economic information. By their nature, contingency tables enable aggregation of potentially sensitive data, limiting disclosure of identifying information. Furthermore, censoring or perturbation can be used to desensitise low cell counts when they arise. However, access to detailed cross-classified tables for research is often restricted by data custodians when too many censored or perturbed cells are required to preserve privacy. In this article, we describe a framework for selecting and combining log-linear models when accessible data is restricted to overlapping marginal contingency tables. The approach is demonstrated through application to housing transition data from the Australian Census Longitudinal Data set provided by the Australian Bureau of Statistics.

Key words: Count data; log-linear model; marginal model; privacy restriction.

1. Introduction

Governments, statistical agencies and data custodians are increasingly using contingency or frequency tables to make data available to the public on a wide range of topics including health, demography, education, and the economy. Tabular data can provide insights on associations among variables and are underpinned by the long-standing statistical framework of log-linear models (see Birch 1963; Bishop et al. 1975; Agresti, 1981; Cameron and Trivedi 1998; Nelder and Wedderburn 1972; Agresti 2002; Bergsma et al. 2009, among others). Importantly, contingency table data can provide statistical outputs whilst preserving the privacy and anonymity of the individuals from which the data are derived.

Making data publicly available is difficult while maintaining the legal and ethical requirements to protect individuals' privacy. Recently, new software and resources have

[1] Queensland University of Technology, School of Mathematical Sciences, GPO Box 2434, Brisbane, Queensland, 4001, Australia. Email: joshuajbon@gmail.com
[2] Australian National University, School of Demography, 9 Fellows Road, Acton, ACT 2601, Australia. Email: bernard.baffour@anu.edu.au
[3] Australian Catholic University, Institute for Learning Sciences and Teacher Education, 229 Elizabeth St, Brisbane, Queensland, 4000, Australia. Emails: Melanie.Spallek@acu.edu.au and Michele.Haynes@acu.edu.au
Acknowledgments: This research was part funded by the Australian Research Council Centre of Excellence for Children and Families over the Life Course (CE140100027). This article uses data from the 2006–2011 Australian Census Longitudinal Data set (ACLD) made available by the Australian Bureau of Statistics (ABS). The authors would like to thank the ABS for their encouragement and helpful comments on this research. The findings and views reported in this article are those of the authors and should not be attributed to the ABS. This work was also supported by resources provided by the Pawsey Supercomputing Centre with funding from the Australian Government and the Government of Western Australia.

enabled organisations to provide safe online access to sensitive data by generating contingency table summaries dynamically from user queries, for example TableBuilder used by the Australian Bureau of Statistics, ABS (ABS 2012). The derived tables are only released after balancing the utility and confidentiality risk (Chipperfield et al. 2016). These *dynamically-generated contingency tables* are a powerful resource for applied researchers to utilise for discovering patterns and associations in the data whilst preserving privacy. In particular, dynamically-generated tables have found favour among national statistical agencies, including the United States Census Bureau, the United Kingdom's Office for National Statistics, Statistics Netherlands, and the Australian Bureau of Statistics (Duncan et al. 2011; Chipperfield et al. 2016), and are often used to release census data.

To mitigate privacy risks, data custodians can employ a number of statistical techniques to control disclosure. For example, they may limit the number of variables allowed to be reported simultaneously, or place limits on the frequency of small cell sizes. Query restrictions, such as these, reduce disclosure risk for sensitive information, but do come at a cost to statistical analysis (Domingo-Ferrer and Mateo-Sanz 1999). Specifically, these restrictions may prohibit the user from accessing all the variables of interest in a single contingency table which is problematic for robust analysis. Practically, users can mitigate these restrictions by requesting a set of separate but overlapping contingency tables to analyse individually. In this article we focus on overlapping tables, specifically a set of contingency tables where the pairwise intersection of variables in any two tables is a common nonempty subset of the available variables. Section 2 provides an illustrative example.

To address privacy and disclosure concerns in the analysis of unit record administrative data, Lee et al. (2017) have recently proposed a modelling framework that computes sufficient statistics from separate data sources that may include subsets of "similar structure" (e.g., subsetting by natural spatial groupings such as state) from a single big database, and potentially subsets from other databases that are relevant to different levels of a hierarchical model. The sufficient statistics are computed by the data custodian, but are combined by the researcher for construction of the log-likelihood to obtain model estimates that approximate those that would be estimated from the full data set. It is further proposed that this modelling framework could be incorporated into data extraction tools provided by data custodians, such as TableBuilder. However, until this has been achieved, researchers will need to rely on analysis of aggregated data from overlapping contingency tables for many applications. In this article, we outline an approach for model estimation by combining output from separate contingency tables.

In the framework of log-linear models we show that, after an appropriate adjustment, model selection can be used to compare overlapping contingency tables, thereby computing relative importance of the explanatory variables. In addition, we re-purpose an existing technique to combine models for the overlapping tables to form the appropriate higher order model allowable by the restricted data.

The article is structured as follows. In Section 2, an illustrative example of a relevant scenario requiring access to and analysis of confidential data is explored. Section 3 provides an overview on (marginal) log-linear models, and describes the methods we use to compare and combine the models in detail. Section 4 applies the methodology to Australian housing tenure transition data from the Australian Longitudinal Census Dataset, and a summary of the method and conclusions is discussed in Section 5.

2. Illustrative Example

Scenario: A regional subset of a population census classifies each person by four sensitive categorical variables. The data set is held securely by a data custodian who has chosen to release the data online using dynamically generated contingency tables. However, the custodian has deemed that releasing the full contingency table (the *super-table*) poses a privacy risk due to the small regional population size (in reality this assessment can be done in real-time based on some measured sparsity of the table requested, see Chipperfield et al. (2016) for example). As such, they will only allow tables with up to two variables (the *marginal tables*) to be released. Under this restriction, there are two analyses that may be of interest – but currently unavailable to researchers. The first involves investigating which two variables (of the four) best explain the count data observed. Meanwhile, the second builds a model that encompasses all four marginal (overlapping) tables.

The above scenario is simplified, but in essence demonstrates the problem this article addresses. The relationship between super- and marginal tables is illustrated in Figure 1, where an inaccessible contingency table (the super-table S) is marginalised into three unique tables (the marginal tables M_1, M_2, and M_3) each with fewer variables than the super-table. In this example, the super-table has four categorical variables, C_1, C_2, C_3, and C_4. All of which can each take one of two values in this example. For simplicity, we label these values with integers 1 and 2. The marginal tables each contain two of the variables from the super-table, always the overlapping variable C_1 and one remaining variable from $\{C_2, C_3, C_4\}$. The count data are aggregated according to the marginalisation of the variables excluded in each marginal table.

The issue here is that the log-linear models are not directly comparable when they are estimated from the marginal tables. Specifically, straight-forward comparison requires that the cell probabilities in the super-table are estimated under the constraints imposed by each marginal model (Bergsma et al. 2009). As the super-table is inaccessible, neither the cell probabilities nor the estimated model parameters can be compared without some adjustment. After addressing this first issue, we will discuss how to perform joint inference on the marginal tables.

3. Methods for Overlapping Marginal Log-Linear Models

3.1. Background

Contingency table cell counts can be used to fit log-linear models formulated under the generalised linear modelling (GLM) framework (Nelder and Wedderburn 1972) with a log link function and Poisson distributed counts. Contingency tables can also be modelled with a multinomial distribution. These are also referred to as log-linear models as the Poisson and multinomial regressions have equivalent maximum likelihood point estimates under mild assumptions (Lang 1996).

The sufficient statistics of marginal log-linear models are the maximum likelihood estimates of the expected frequencies under the corresponding marginal contingency table. This follows from Birch's theorem (Birch 1963), which implies that the maximum likelihood estimates match the marginal distributions and also ensures that the associations and interactions satisfy the model-implied patterns. In other words, there is a unique set of

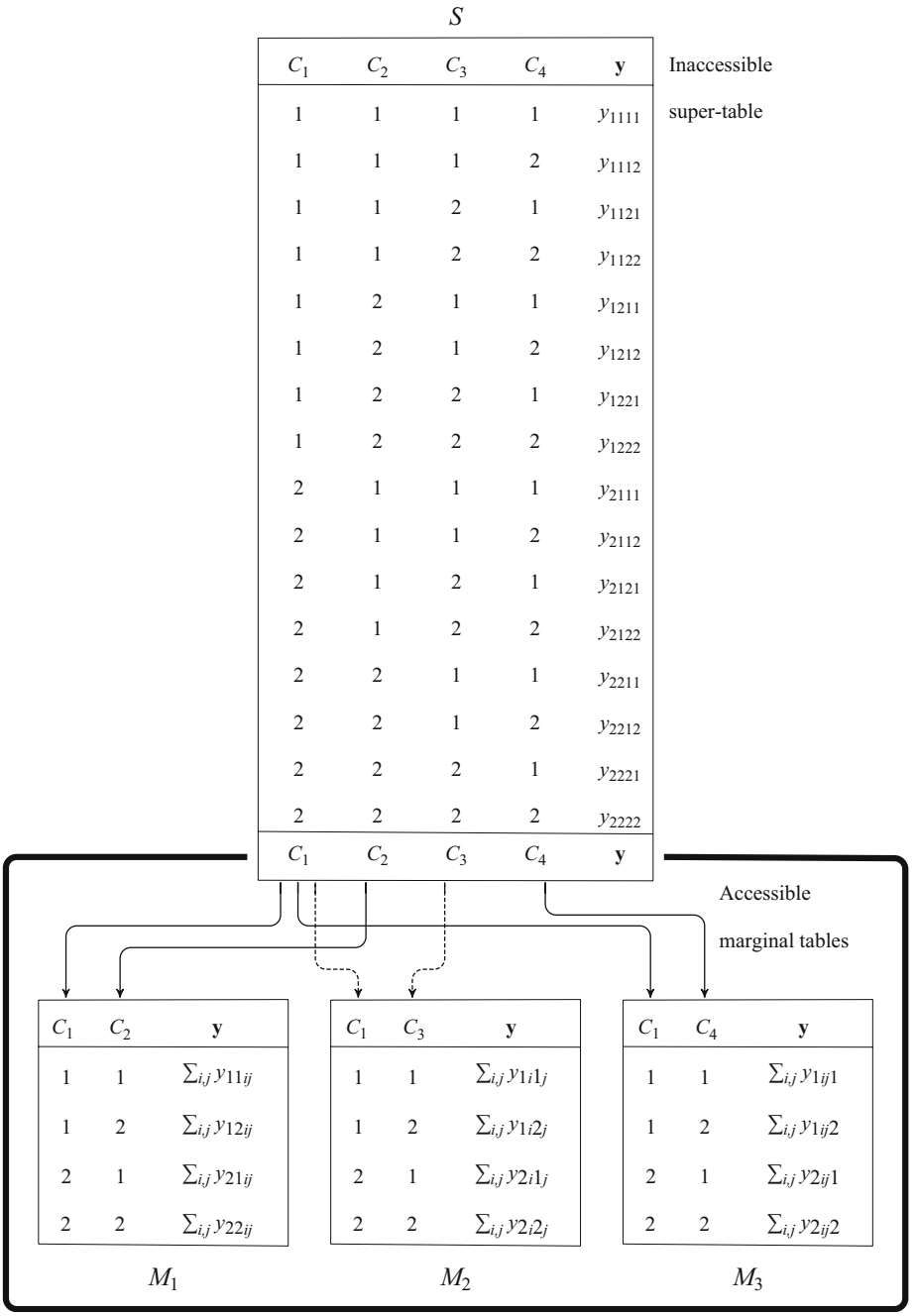


Fig. 1. Illustration of accessible marginal tables nested in an inaccessible super contingency table. Each marginal table contains C_1, the overlapping variable, and one of the remaining variables from the super-table $\{C_2, C_3, C_4\}$.

fitted values that both satisfy the marginal model and match the data in the sufficient statistics, and this unique solution is the maximum likelihood estimate. Regression coefficients from log-linear regression models can be equivalently specified as associations, expected counts or cell probabilities.

Following notation from Lang (1996), a probability vector $\mathbf{p}$ containing probabilities from a contingency table can be specified by the log-linear model

$$\log(\mathrm{E}(\mathbf{p})) = \boldsymbol{\xi} = \mathbf{X}\boldsymbol{\beta} \tag{1}$$

where the cell probabilities are related to the cell counts $\boldsymbol{\mu}$ by $\mathbf{p} = n^{-1}\boldsymbol{\mu}$, and n is the sum of all counts. The vector $\boldsymbol{\xi}$ contains the expected probabilities on the log-scale, and $X\boldsymbol{\beta}$ codes the associations between cells as in generalised linear modelling with a Poisson distribution.

Marginal log-linear models have been studied extensively (see Bergsma and Rudas 2002; Bergsma et al. 2009, and references therein). These can be used to fit models where some associations (or equivalently cell probabilities) among the table cells are restricted or removed. Marginal models can be estimated with restrictions specified by Lagrangian multipliers added to the log-likelihood of the full model. Hence, the full data set is used in the fitting procedure. This is appealing since it ensures that when several marginal models are estimated, they are comparable because only the constraints have changed (the likelihood is still based on the same data set). While this study considers marginal models, the issue addressed here differs with regard to availability of the data for analysis. Specifically, the joint (or full) contingency table is not accessible, hence the constrained estimation approach to marginal models is not possible. We consider the situation where several marginal, overlapping contingency tables are accessible instead.

To clearly distinguish the general contingency tables we reference within this article, we refer to the inaccessible table which would encompass all of the marginal tables as the *super-table*. In particular, we are interested in fitting and comparing decomposable graphical log-linear models on this super-table.

Decomposable graphical models have several advantages over their non-decomposable counterparts. First, the maximum likelihood estimates can be found explicitly. Second, closed form solutions exist for the sufficient statistics. Third, a necessary condition for decomposability is that the models are hierarchical; the absence of an interaction forces all related higher-order interactions to be excluded, which aids in interpretability. Finally, an attractive feature of decomposable graphical models is that they can be interpreted in terms of their patterns of conditional independencies, which can also be displayed graphically.

A decomposition method for fitting hierarchical log-linear models with large contingency tables was proposed by Dahinden et al. (2010). In essence, associated subsets of variables are identified from a super-table, after which cell probabilities for each subset (or marginal table) are estimated using log-linear models. The results are combined using the decomposability property of graphical models (Lauritzen 1996). Sparsity can be considered using Lasso or model selection on the sub-models. In our application, with access restricted to marginal tables, it is the decomposability property that can be used to combine the results from several marginal tables. This approach is described in Subsection 3.3.

3.2. Comparing Models from Overlapping Tables

When the super-table is inaccessible, the constrained formulation of marginal models is not possible to implement. As such, the marginal models must be estimated from the different marginal data sets available, as illustrated in Section 2.

The approach addressed in this article is the converse of the situation in Allison (1980), where the results demonstrated the equivalence in estimated probabilities between collapsed and uncollapsed data sets. Their intention was to fit marginal models on contingency tables while avoiding collapsing the table itself. We, on the other hand, would like to estimate the same probabilities (or equivalently frequencies) for the full table, using only the collapsed data set.

Specifically, Allison (1980) demonstrates how the estimated cell frequencies from the full table and the collapsed table are equivalent when the frequencies from the former are also collapsed. The two frequency vectors share the same association structure (model equation) that must be collapsible for the given data. Collapsibility, as discussed in Bishop et al. (1975), is the key to ensuring these frequencies are equivalent (after adjusting for multiplicity) – it says that collapsing over one set of variables will not affect the parameters in a second set, if the two sets are independent. In our case, and in Allison (1980), the collapsed variables are not included in the model, and are therefore independent of the variables that are included.

In our analysis we fit several Poisson GLMs with overlapping explanatory categories with their respective collapsed or marginal data. We adjust the log-likelihood of each marginal model (after estimation) so that it is as if each model had been estimated using the super-table data, which contains all categorical variables used in every model. The association structure of each marginal model does not change. We must make this adjustment so that model selection techniques can be appropriately applied. As mentioned, the adjustment relies on the equivalence between probabilities from the model estimated with marginal data to the same model fitted using the super-table (Bishop et al. 1975; Allison 1980). The linear model coefficients (describing the associations) will be equal under both scenarios, except for the intercept terms which differ depending on the number of rows in each model matrix.

Below, we derive the exact adjustment needed to compare overlapping marginal models using likelihood-based metrics such as AIC (Akaike 1974). This adjustment is implicitly linked to sufficiency in marginal log-linear models and the constrained formulation of marginal log-linear models (Bergsma et al. 2009), but the authors have not been able to locate a previous derivation in the literature. Note that in the following derivation we describe two estimated vectors of probabilities that share a single association structure, the *model equation*. The first model is hypothetically estimated using the super-table as data, since this data is unavailable in our application, while the second model uses marginal data sufficient to estimate the model equation of interest.

In order to prove our result, we start by establishing a connection between two log-linear models estimating the model equation (identical design matrix, $\mathbf{X}$, and coefficient vector, $\boldsymbol{\beta}$). The models differ only in the data used to fit each of them – but both data sets are sufficient to estimate the given association structure. The first model uses the vector of

counts $\mathbf{y} = \left[y_1\, y_2 \cdots y_m\right]^\top$, of length m (counts from the super-table), while the second uses a collapsed vector of counts $\mathbf{y}^\kappa = \left[y_1^\kappa\, y_2^\kappa \cdots y_{m^\kappa}^\kappa\right]^\top$, of length m^κ. This technique is then applied to the set of marginal, overlapping data sets, so that correct model comparisons can be made.

The following assumptions are required in order to equate the probabilities estimated from the marginal data to those estimated using the super-table:

1. The association structures, or model equations, to be estimated for each data set of counts ($\mathbf{y}$ and $\mathbf{y}^\kappa$) are identical.
2. The marginal data ($\mathbf{y}^\kappa$) is sufficient to estimate the model equation.
3. The counts (or cells) from the super-table, $\mathbf{y}$, have been collapsed to $\mathbf{y}^\kappa$ using $\mathbf{M}$, as described in Equation (2).
4. The variables that are collapsed are irrelevant under the given model equation.

Of the above assumptions the fourth is the strongest, although it is one that we have to make under any modelling strategy when only the marginal contingency table is available.

Let $\mathbf{M}$ be a matrix that collapses a vector of counts, $\mathbf{y}$, from the super-table to the observed counts in the marginal table, $\mathbf{y}^\kappa$. Specifically, these two vectors are related by

$$\mathbf{y}^\kappa = \mathbf{M}\mathbf{y}. \tag{2}$$

The matrix $\mathbf{M}$ is a $m^\kappa \times m$ matrix containing only zeros and ones. Every column of $\mathbf{M}$ contains only one unit element, while every row contains $r = m/m^\kappa$ (an integer) unit entries. The matrix $\mathbf{M}$ is a type of incidence matrix that sums the counts in $\mathbf{y}$ to the counts in $\mathbf{y}^\kappa$ and describes the marginalisation of the model. The following identity is useful:

$$y_i^\kappa = \sum_{j=1}^{m} M_{ij} y_j = \sum_{j:M_{ij}=1} y_j \tag{3}$$

where the first equality holds by definition, and the last equality holds since each element of M is either one or zero.

Using the illustrative scenario in Section 2 as an example, the matrix $\mathbf{M}$ that collapses the counts in super-table S to the marginal table M_1 is

$$\mathbf{M} = \begin{bmatrix} \mathbf{1}_4^\top & & & \\ & \mathbf{1}_4^\top & & \\ & & \mathbf{1}_4^\top & \\ & & & \mathbf{1}_4^\top \end{bmatrix} \quad \text{where } \mathbf{1}_4^\top = \left[1\ 1\ 1\ 1\right]$$

with zeros filling the blank cells.

The estimated cell probabilities from the model with marginal data, say $\hat{\mathbf{p}}^\kappa$, are related to the cell probabilities, $\hat{\mathbf{p}}$, in the constrained marginal formulation by

$$\hat{\mathbf{p}}^\kappa = \mathbf{M}\hat{\mathbf{p}} \tag{4}$$

under Assumptions 1–4. In other words, the estimated probabilities of the models with collapsed data and uncollapsed data (super-table) are equal, up to a multiplicative constant that accounts for the difference in the number of cells for each data set. This is only true because the association structure being fitted, as in Allison (1980), does not change across the marginal (collapsed) data and super-table (uncollapsed data).

Another property of the estimated probabilities, $\hat{\mathbf{p}}$, is that its elements repeat according to the pattern of zeros and ones in $\mathbf{M}$, as such

$$\hat{p}_s = \hat{p}_t \text{ if } M_{is} = M_{it} = 1, \text{ for some } i \in \{1, 2, \ldots, m^{\kappa}\}. \tag{5}$$

That is, cells in the super-table that share the same association structure under the marginal model (the marginal model mean equation) will have the same estimated probability (Allison 1980).

The identity for the counts in Equation (3), also holds for the probabilities, that is

$$\hat{p}_i^{\kappa} = \sum_{j:M_{ij}=1} \hat{p}_j \tag{6}$$

which can be combined with Equation (5) to give

$$\hat{p}_i^{\kappa} = r\hat{p}_j \text{ if } M_{ij} = 1, \tag{7}$$

since each $\hat{p}_j$ in the sum of Equation (6) for a given i are equal, and there are r probabilities being summed in each row of $\mathbf{M}$. To explain intuitively, as per Equation (5), using the super-table to estimate the marginal model (instead of a sufficient collapsed data set) results in dividing each probability evenly across r cells of the super-table since the same mean equation is repeated r times. It is also true that

$$\sum_{j:M_{ij}=1} 1 = r, \tag{8}$$

that is the row sums of M are equal to r, the multiplicity factor that arises from collapsing the cells and hence probabilities.

Using the estimated cell probabilities, the Poisson log-likelihood for the model with collapsed, or marginal data is

$$\log L\left(\hat{\mathbf{p}}^{\kappa}|\mathbf{y}^{\kappa}\right) = \sum_{i=1}^{m^{\kappa}} \left(y_i^{\kappa}\left(\log \hat{p}_i^{\kappa} + \log n\right) - n\hat{p}_i^{\kappa} - \log y_i^{\kappa}!\right) \tag{9}$$

for vector of observed count data, $\mathbf{y}^{\kappa}$ where $n = \sum_{i=1}^{m^{\kappa}} y_i^{\kappa}$. The log-likelihood for the model with marginal data can be rewritten using Equations (3), (7), and (8) in the following way

$$
\begin{aligned}
&\log L\left(\hat{\mathbf{p}}^{\kappa}|\mathbf{y}^{\kappa}\right) \\
&\quad = \sum_{i=1}^{m^{\kappa}} \left(\left(\sum_{j:M_{ij}=1} y_j \right) \left(\log\left(\hat{p}_i^{\kappa}\right) + \log n \right) - n\hat{p}_i^{\kappa} \right) + c\left(\mathbf{y}^{\kappa}\right) \\
&\quad = \left(\sum_{i=1}^{m^{\kappa}} \left(\sum_{j:M_{ij}=1} y_j \left(\log\left(\hat{p}_i^{\kappa}\right) + \log n \right) \right) - \frac{1}{r} \sum_{j:M_{ij}=1} n\hat{p}_i^{\kappa} \right) + c\left(\mathbf{y}^{\kappa}\right) \\
&\quad = \left(\sum_{i=1}^{m^{\kappa}} \sum_{j:M_{ij}=1} \left(y_j \left(\log\left(\hat{p}_i^{\kappa}\right) + \log n \right) - \frac{n}{r}\hat{p}_i^{\kappa} \right) \right) + c\left(\mathbf{y}^{\kappa}\right) \\
&\quad = \left(\sum_{i=1}^{m^{\kappa}} \sum_{j:M_{ij}=1} \left(y_j \left(\log\left(r\hat{p}_j\right) + \log n \right) - n\hat{p}_j \right) \right) + c\left(\mathbf{y}^{\kappa}\right) \\
&\quad = \sum_{j=1}^{m} \left(y_j \left(\log \hat{p}_j + \log r + \log n \right) - n\hat{p}_j \right) + c\left(\mathbf{y}^{\kappa}\right) \\
&\quad = \log L\left(\hat{\mathbf{p}}|\mathbf{y}\right) - c\left(\mathbf{y}\right) + n \log r + c\left(\mathbf{y}^{\kappa}\right).
\end{aligned}
$$

The integer n is the total of the counts, $n = \sum_{i=1}^{m} y_i = \sum_{j=1}^{m^{\kappa}} y_j^{\kappa}$, whose equality across holds approximately when perturbations have been added for further privacy. The constants are defined as $c\left(\mathbf{y}^{\kappa}\right) = -\sum_{i=1}^{m^{\kappa}} \log y_i^{\kappa}!$ and $c\left(\mathbf{y}\right) = -\sum_{j=1}^{m^{\kappa}} \log y_i!$. Thus an equivalence between the log-likelihood using the super-table, $\log L\left(\hat{\mathbf{p}}|\mathbf{y}\right)$, and log-likelihood using the marginal data, $\log L\left(\hat{\mathbf{p}}^{\kappa}|\mathbf{y}^{\kappa}\right)$, can be expressed as

$$
\log L\left(\hat{\mathbf{p}}|\mathbf{y}\right) - c\left(\mathbf{y}\right) = \log L\left(\hat{\mathbf{p}}^{\kappa}|\mathbf{y}^{\kappa}\right) - c\left(\mathbf{y}^{\kappa}\right) + n\left(\log m^{\kappa} - \log m\right). \tag{10}
$$

The constant $c(\mathbf{y})$ cannot be calculated because the super-table is inaccessible. However, for quantities where a difference is of interest, such as information criteria, the $c(\mathbf{y})$ cancel out. The relative adjusted AIC for a marginal-data model with probability vector $\mathbf{p}^{\kappa}$ and size m^{κ} can be calculated as

$$
\mathrm{aAIC}\left(\mathbf{p}^{\kappa},\mathbf{y}^{\kappa},k,m^{\kappa}\right) = 2k - 2\left(\log L\left(\hat{\mathbf{p}}^{\kappa}|\mathbf{y}^{\kappa}\right) - c\left(\mathbf{y}^{\kappa}\right) + \left(\log m^{\kappa} - \log m\right) \sum_{i=1}^{\kappa} y_i^{\kappa} \right). \tag{11}
$$

The above aAIC is relative because it does not include the constant from the log-likelihood. The number of association parameters is k, and the constant m should be fixed for a given set of overlapping marginal tables. It is the product of the number of levels in the set of unique variables among all marginal tables (see Subsection 4.2).

3.3. Combining Models for Overlapping Tables

We refer to the combination of marginal models as *stitching* and refer to the result as a *stitched* model. The stitching process takes the several overlapping log-linear models and generates the equivalent model if all parameters had been estimated jointly with a contingency table that would enable this. The method we describe has been re-purposed from Dahinden et al. (2010) who consider the case where the full data set is available.

The stitching of marginal models is possible when the set of marginal models to be combined together are a decomposition of a possible hierarchical model on the super-table. Decomposability can be described by considering the log-linear regression as a graphical model (Darroch et al. 1980) with graph $G = (V, E)$, having vertex set V, and edge set E. Define a subgraph of G induced by $W \subset V$ as $G[W] = (W, \{(u, v) \in E : u, v \in W\})$, effectively the graph remaining from G after removing all vertices absent from W (and all hanging edges). A partition of the vertex set, V, into $\{A, S, B\}$ is a decomposition if $G[S]$ separates $G[A]$ from $G[B]$, and $G[S]$ is a complete graph. A vertex subset $S \subset V$ is a (vertex) separator for A and B if its removal from G separates A and B into disconnected components. We refer to the vertex set S as the separator.

The decomposability of a graph is defined recursively; a graph is decomposable if it is complete or if there exists a decomposition $\{A, S, B\}$ such that the subgraphs $G[A \cup S]$ and $G[S \cup B]$ are decomposable. We refer the reader to Leimer (1993) for further discussion. An example decomposable graph is shown in Figure 2 and a comprehensive guide can be found in Darroch et al. (1980).

If a graph is decomposable then a relationship exists between the full graph and its complete subgraphs: the separators and cliques (vertex set W is a clique if $G[W]$ is a complete graph and, for our purposes, not a separator) (Frydenberg and Lauritzen 1989).

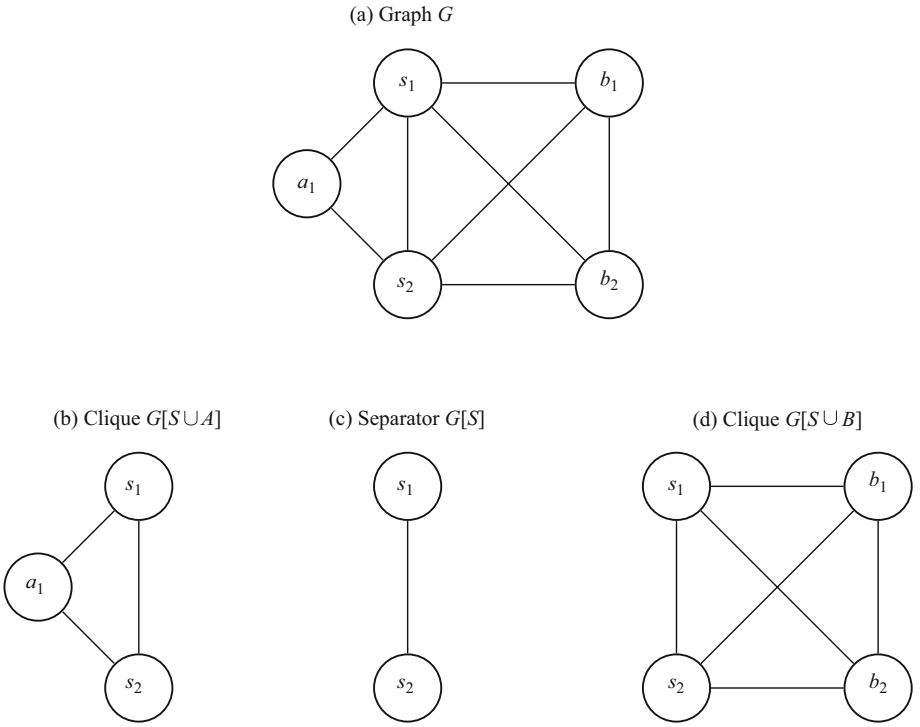


Fig. 2. (a) Example of decomposable graph, G, with separator $S = \{s_1, s_2\}$. After the removal of S from the graph, the subgraphs with nodes $A = \{a_1\}$ and $B = \{b_1, b_2\}$ are disconnected. (b) Clique graph $G[S \cup A]$. (c) Separator graph $G[S]$. (d) Clique graph $G[S \cup B]$. Notice that $G[S \cup A]$ and $G[S \cup B]$ are complete, and hence this graph satisfies the decomposability definition. Moreover, if $G[S \cup B]$ were not complete, it would need to be decomposable to satisfy the recursive definition of decomposability.

The relationship is the special structure afforded the graph — conditioning on the separator vertices emits a conditional decomposition of the graph. In the case of statistical models where each vertex has associated parameter(s), separators act as the only intermediaries between cliques, and fixing their values results in independence between the remaining cliques (Frydenberg 1990). Let the set of cliques be $\mathcal{C}$, and set of separators be $\mathcal{S}$ and note that these sets can be constructed using the recursive definition of decomposability above.

Following Dahinden et al. (2010) we change notation slightly to accomodate the graph theory used in this section. Let $p(i)$ be the probability of belonging to particular categories of a set of variables denoted by i, from a decomposable graph (log-linear model). Let $p(i_C)$ and $p(i_S)$ denote the probability of the same categories but from the clique C and separator S respectively. In terms of log-linear modelling, $p(i)$ is the estimated probability of a particular observation from the super-table, whereas $p(i_C)$ and $p(i_S)$ are calculated from specific marginal models (derived from the overlapping tables we have access to). The relationship in logarithmic terms is

$$\log p(i) = \sum_{C \in \mathcal{C}} \log p(i_C) - \sum_{S \in \mathcal{S}} v(S) \log p(i_S) \tag{12}$$

where $v(S)$ is the index of separator S, describing the number of times S acts as a separator (Lauritzen 1996). Using the relation in Equation (12), the estimated marginal models can be stitched or combined together. The resulting probabilities are from the equivalent joint model on the inaccessible super-table. Equation (12) accounts for the multiplicity of the separator in the estimates from the cliques in the decomposed graph. For example, the separator S appears in both cliques shown in Figure 2. In this case, the index of the separator is $v(S) = 2 - 1 = 1$ (see (Lauritzen (1996) for further details).

In the case of a log-linear model with a non-decomposable graphical counterpart, a minimal triangulation can be used in order to form a decomposable graph [see Rose et al. 1976; Olesen and Madsen 2002, for example]. Under the overlapping structure we consider, the graph generated by stitching saturated marginal models together is already decomposable, so no triangulation is needed. However, for stitching non-saturated marginal models together we suggest beginning with the triangulation equivalent to the saturated models, then using thinning (removing edges added during triangulation) to construct a minimal triangulation (Jones and Didelez 2017). In some circumstances, edge removal (removing existing model edges) may also be necessary to guarantee a model that is both decomposable and graphical – in order to ensure that the necessary sufficient statistics are available from the marginal models.

Our analysis in Section 4 stitches the saturated models from each of the marginal tables together to form a joint model across all available combinations of variables. The estimated probabilities from the resulting model can be used to calculate standard model summaries, such as estimated association coefficients, prevalence ratios, and information criteria.

4. Analysing Housing Transitions from the ACLD

Purchasing a home in Australia is a significant stage in an individual's life course and the prevalence of home ownership is an important indicator of a country's economic

performance. Understanding the drivers of transitions into home ownership is therefore of considerable social and economic interest and is often the subject of life course research (see, for example Spallek et al. 2014). In Australia, a rich source of data on housing tenure transitions is the Australian Census of Population and Housing ("Census"), conducted by the ABS. We investigate home ownership transitions and their associations with demographic factors using a derivative of the 2006 and 2011 Censuses, the Australian Census Longitudinal Dataset (ACLD), (Chipperfield et al. 2017).

4.1. Data

The 2006–2011 ACLD contains information from a five-percent random sample of the Australian population selected from the 2006 Census and then subsequently linked to the 2011 Census. The final linked data set (the ACLD) consists of 800,759 records (ABS 2013). The differences between the original sample of the 2006 Census and the final linked sample are attributable to either deaths and overseas departures that occurred between the 2006 and 2011 censuses, or due to unsuccessful linkages because of inconsistent or missing information. The ACLD may be accessed with the TableBuilder software product, an online table creator that allows users to build contingency tables from ABS data without accessing unit records (Chipperfield et al. 2016; ABS 2012). TableBuilder is subject to both query restrictions and perturbations to ensure anonymity of the individuals from the underlying data.

The ideal approach to investigating home ownership transitions using the ACLD would be to create a super-table including housing tenure transitions cross-tabulated with all other variables of interest. However, due to query restrictions, requests to TableBuilder with more than 14 variables and 44 categories exceed the cell limit allowed for contingency tables. Therefore, a new strategy is needed to develop a model with the required variables. We created what we refer to as a base contingency table including housing tenure transitions between 2006 and 2011, categorised by age, and gender. Transitions of each variable, previously shown to be associated with housing tenure transitions, were added to the base contingency table to form a new, separate contingency table. This resulted in six contingency tables (CT1–CT6), where CT1-base contains age, gender and housing tenure transitions; CT2-children contains all CT1-base variables and children status transition; CT3-family contains all CT1-base variables and family status transition; CT4-labour contains all CT1-base variables and labour transition; CT5-marital contains all CT1-base variables and marital transition; and CT6-geography contains all CT1-base variables and geographical transition. To assess which of these variables was most strongly associated with housing tenure transitions, we applied a set of log-linear models to each of the contingency tables CT1–CT6 and used the AIC to select the best model (BM1–BM6) for each contingency table (CT1–CT6). The set of log-linear models applied to each of CT1–CT6 ranged from models with single main effects to saturated models which includes all interactions. After adjusting the AIC for each best model (BM1–BM6), as discussed in Section 3, we compared BM1–BM6 to identify which of the demographic variables has the strongest association with housing tenure transitions in conjunction with age group and sex.

The analytical sample in this example is restricted to non-Indigenous Australians aged between 20 and 60 years old who did not own a house outright in 2006. We do not consider

those aged over 60 years old, because individuals in this age group experience transitions in home ownership related to different events, for example retirement. Our final sample consists of 260,595 individuals from the ACLD who have data linked between the two census time points.

Each of the six contingency tables (CT1–CT6) contain variables age, sex, housing tenure transition, and one additional transition variable. Housing tenure transitions are distinguished between renting (or other) and owning with a mortgage, coded as 1 and 2 respectively. Individuals owning a home outright are excluded from this analysis. For illustrative analysis we consider the core variables of interest to be sex (coded as 1 = 'Male'; 2 = 'Female') and age (coded as 1 = '21 to 30 years old'; 2 = '31 to 40 years old'; and 3 = '41 to 60 years old'), While the additional transition variables are; children status (coded as 1 = 'No children (0–4 years)'; 2 = 'One child (0–4 years)'; 3 = 'Two or more children (0–4 years)'; and 4 = 'Not applicable'), family status (coded as 1 = 'Couple family, no children'; 2 = 'Couple family, children under 15'; 3 = 'Couple family, no children under 15'; 4 = 'One parent family, children under 15'; 5 = 'One parent family, no children under 15'; 6 = 'Other family, or not applicable'), labour status (coded as 1 = 'Employed'; 2 = 'Unemployed'; 3 = 'Not in the labour force or other'), marital status (coded as 1 = 'Married'; 2 = 'Never married'; and 3 = 'Separated, divorced, or widowed'), and geography status (coded as 1 = 'Australia Major Cities'; 2 = 'Australia Regional'; and 3 = 'Remote or other'). Table 1 summarises the demographic variables of the analytical sample from the ACLD.

The analytical sample includes slightly more females than males (52.2% versus 47.8%), with the majority being aged between 41–60 years (40.7%). In 2006, 36.6% of individuals were renting, which decreased to 31.5% in 2011. The percentage of people owning their own home with a mortgage increased, as would be expected with the same ageing population, from 63.4% to 68.5% during these five years to 2011. In 2006 the majority of people had no children between the age 0–4 years (41.1%), defined their family status as being a couple family with children less than 15 years old (40.0%), were employed (80.0%), married (55.9%) and lived in an Australian major city (72.3%). With regard to these variables and related categories, there were no major differences in their distributions between 2006 and 2011 with the exception of the percentage of married individuals, which increased by 5.5%. While the distribution of individuals across the categories of most variables was stable, this masks the changes at the individual level.

4.2. Empirical Validation

Before conducting the main analysis, we validate the comparison method discussed in Subsection 3.2 by fitting the identical models on the six contingency tables (CT1–CT6) generated by TableBuilder and introduced in Subsection 4.1. The "validating" log-linear model regresses the counts from each table with combinations of the categories in Table 1 from sex (S), age (A), and housing tenure transition ($T_{2006} \times T_{2011}$). The notation $T_{2006} \times T_{2011}$ represents the interactions of the categories in housing tenure in 2006 and 2011, namely, rent to rent, rent to own, own to rent, own to own (ownership is always with a mortgage). In general, we will suppress multiplication symbol in $C_1 \times C_2$ and simply write the $C_1 C_2$. The validating model can therefore be represented by the notation $S A T_{2006} T_{2011}$.

Table 1. Categories of variables taken from Australian census longitudinal data set with aggregated counts and percentages.

Variable	Variable name	Value	Value name	Counts (percent) 2006	(%)	2011	(%)
S	Sex	1	Male	124,622	(47.8)		
		2	Female	135,973	(52.2)		
A	Age bracket	1	21 to 30 years old	69,312	(26.6)		
		2	31 to 40 years old	85,237	(32.7)		
		3	41 to 60 years old	106,046	(40.7)		
T	Housing tenure status	1	Rent (or other)	95,435	(36.6)	82,205	(31.5)
		2	Owned (mortgage)	165,160	(63.4)	178,390	(68.5)
C	Children status	1	No children (0–4 years)	107,175	(41.1)	115,056	(44.2)
		2	One child (0–4 years)	39,469	(15.1)	36,039	(13.8)
		3	Two or more children (0–4 years)	16,587	(6.4)	16,363	(6.3)
		4	Not applicable	97,323	(37.3)	93,096	(35.7)
F	Family status	1	Couple family, no children	53,473	(20.5)	53,026	(20.3)
		2	Couple family, children under 15	104,238	(40.0)	103,209	(39.6)
		3	Couple family, no children under 15	32,218	(12.4)	37,941	(14.6)
		4	One parent family, children under 15	13,715	(5.3)	12,570	(4.8)
		5	One parent family, no children under 15	10,116	(3.9)	11,690	(4.5)
		6	Other family, or not applicable	46,780	(18.0)	42,104	(16.2)
L	Labour status	1	Employed	208,491	(80.0)	210,268	(80.7)
		2	Unemployed	8,630	(3.3)	7,529	(2.9)
		3	Not in the labour force, or other	43,466	(16.7)	42,790	(16.4)
M	Marital status	1	Married	145,543	(55.9)	159,959	(61.4)
		2	Never married	80,796	(31.0)	61,016	(23.4)
		3	Separated, divorced, or widowed	34,197	(13.1)	39,561	(15.2)
G	Geography status	1	Australia Major Cities	188,400	(72.3)	189,022	(72.5)
		2	Australia Regional	68,204	(26.2)	67,799	(26.0)
		3	Remote or other	3,964	(1.5)	3,747	(1.4)

Table 2. Adjusted AIC of marginal contingency tables on validating model: S A T_{2006} T_{2011}.

Table	unadjusted AIC	n. cells	Relative adjusted AIC	
			aAIC − min(aAIC)	$\frac{\text{aAIC}}{\text{min(aAIC)}} - 1$ %
CT1-Base	308	24	428	0.019
CT2-Children	450,743	384	140	0.006
CT3-Family	632,602	864	16	0.001
CT4-Labour	630,264	216	327	0.014
CT5-Marital	525,031	216	0	0.000
CT6-Geography	682,548	216	263	0.012

To use the adjusted log-likelihood (and hence the adjusted AIC) in Equation (10) we need to calculate the size of the super-table, m, which would have CT1–CT6 as marginal models. This can be calculated by taking all the unique variables in CT1–CT6, then finding the product of the number of levels for each variable. Alternatively, since CT1 is nested in each marginal model m can be calculated by $m = m_1^\kappa \prod_{i=2}^{6} \left(m_i^\kappa / m_1^\kappa\right)$ where m_i^κ is the number of cells in CTi.

Table 2 contains the unadjusted and adjusted AIC values for each table using the model $S\ A\ T_{2006}\ T_{2011}$. The unadjusted AIC values demonstrate that although each contingency table is derived from the same underlying information, and in aggregate will be almost identical, the AIC values from an identical model still differ. There are two reasons why the AIC values are different. Firstly, the number of cells differ across contingency tables (see "n. cells" in Table 2) and secondly, some counts have been perturbed. Since each contingency table is a non-identical data set the justification for the AIC no longer holds. As discussed in Section 3, the relative adjusted AICs can be used for model comparison instead. The relative adjusted AIC is shown as a value and as a proportion in columns 3 and 4 of Table 2. The relative value is given because the AIC from the super-table can only be evaluated up to a constant, as noted in Subsection 3.2. The proportional value of the adjusted AIC is shown to demonstrate the magnitude of the differences of the adjusted AIC values. The adjusted AIC results show that there is a small discrepancy in the adjusted AIC of less than 0.02%. This error is due to the perturbations that differ for every unique data set retrieved from TableBuilder (Chipperfield et al. 2016).

4.3. Comparing Marginal Models from the ACLD

To identify the best model (BM1–BM6) for each of the contingency tables CT1–CT6, we performed step-wise selection using the unadjusted AIC. The analysis was conducted in the statistical language `R` (R Core Team 2016). Following this, the adjusted AIC is calculated for each of the best models (BM1–BM6) so that a comparison is possible across BM1–BM6. Table 3 shows these relative adjusted AIC along with the total number of observations (m^κ), the number of coefficients in the model (k) and remaining degrees of freedom (df). Table 3 is ordered by the relative adjusted AIC. The model including the geographical transition was selected as the best model overall (of BM1–BM6), followed by the models including labour status and family status transitions. Unsurprisingly, the

Table 3. Adjusted AIC comparison of the best models for each marginal contingency table (CT1–CT6).

Best model	m^{κ}	k	df	Relative adjusted AIC	
				aAIC − min(aAIC)	Ranking
BM6-Geography	216	162	54	0	1
BM4-Labour	216	200	16	52,502	2
BM3-Family	864	684	180	54,644	3
BM5-Marital	216	178	38	157,406	4
BM2-Children	384	324	60	233,260	5
BM1-Base	24	22	2	680,874	6

base model which contains only age, sex and housing tenure transition yielded the highest relative adjusted AIC and hence was the worst performing model by the AIC.

Table 4 details the interactions that are present in the best model (BM1–BM6) for each of the six contingency tables (CT1–CT6). Each of BM1–BM6 contains the main effects and all 2nd degree interactions, but inclusion of higher degree interactions differed across models. For example BM1-base includes all 3rd but no 4th degree interactions, namely the model represented by $A\ T_{2006}\ T_{2011} + S\ T_{2006}\ T_{2011} + S\ A\ T_{2011} + S\ A\ T_{2006}$.

Table 4. Best models (BM1–BM6) for each contingency table (CT1–CT6) by step-wise AIC. All BMs have main effects and 2nd degree interactions.

Best model	3rd degree interactions	4th degree interactions	5th degree interactions
BM1-Base	All	None	–
BM2-Children	All	All, excluding: $S\ A\ T_{2011}\ C_{2006}$, $S\ A\ T_{2011}\ C_{2011}$	$S\ T_{2006}\ T_{2011}\ C_{2006}\ C_{2011}$, $S\ A\ T_{2006}\ C_{2006}\ C_{2011}$, $A\ T_{2006}\ T_{2011}\ C_{2006}\ C_{2011}$
BM3-Family	All	All, excluding: $S\ A\ T_{2011}\ F_{2006}$	$S\ T_{2006}\ T_{2011}\ F_{2006}\ F_{2011}$, $A\ T_{2006}\ T_{2011}\ F_{2006}\ F_{2011}$
BM4-Labour	All	All	$S\ A\ T_{2006}\ L_{2006}\ L_{2011}$, $S\ A\ T_{2011}\ L_{2006}\ L_{2011}$, $S\ T_{2006}\ T_{2011}\ L_{2006}\ L_{2011}$, $A\ T_{2006}\ T_{2011}\ L_{2006}\ L_{2011}$
BM5-Marital	All	All, excluding: $S\ A\ T_{2011}\ M_{2006}$ $S\ A\ T_{2006}\ T_{2011}$, $S\ T_{2006}\ T_{2011}\ M_{2011}$	$S\ A\ T_{2006}\ M_{2006}\ M_{2011}$, $A\ T_{2006}\ T_{2011}\ M_{2006}\ M_{20ll}$
BM6-Geography	All, excluding: $S\ T_{2006}\ G_{2006}$, $S\ T_{2011}\ G_{2006}$	All, excluding: $S\ A\ T_{2006}\ G_{2006}$, $S\ A\ T_{2011}\ G_{2006}$, $S\ T_{2006}\ T_{2011}\ G_{2006}$, $S\ T_{2006}\ G_{2006}\ G_{2011}$, $S\ T_{2011}\ G_{2006}\ G_{2011}$	$S\ A\ T_{2006}\ T_{2011}\ G_{2011}$, $A\ T_{2006}\ T_{2011}\ G_{2006}\ G_{2011}$

BM6-geography was found to be the best model including two of a possible six 5th degree interactions, two-thirds of the possible 4th degree interactions, and almost all 3rd degree interactions as specified in Table 4. Note that all models BM1-BM6 except the base model (BM1-base) included a 5th degree interaction containing age, housing tenure transition, and the transition of their additional variable. This indicates that there is an important association with tenure transition across all the additional variables investigated.

Selected prevalence ratios from the BM6-geography are described in Table 5 to demonstrate some inferences that can be draw from this model. The prevalence ratios indicate that individuals remaining in the city (i.e., city to city transition) were most associated with transitioning from renting in 2006 to owning (with a mortgage) in 2011, across all age groups and sexes. The second strongest association was observed for individuals remaining in regional locations, which also persisted across all age groups and sexes. Females, aged 21 to 30 in 2006, and remaining in the city had the highest mean association with transition into home ownership. The prevalence ratio indicated that compared to females, aged 21 to 30, who remain in remote areas, females in the same age-group who remain in the city were approximately 108.54 times more likely to move from renting to owning. Comparing males to females in each age group and geographical transition, shows that the confidence intervals for the prevalence ratios overlap. This is to be expected since the two 5th degree interactions relevant to the best geography model (see Table 4) do not include an interaction between sex and the geographical locations in both years. Some of these results are not supported by the analysis once we combine the models in Subsection 4.4.

4.4. Combining the ACLD Marginal Models

Table 6 contains a selection of results from the stitched model; the model combined from five marginal models using the decomposability property discussed in Subsection 3.3. It shows the same values as Table 5, but for the stitched model which has the (factored) model equation

$$SAT_{2006}T_{2011}[C_{2006}C_{2011}+F_{2006}F_{2011}+L_{2006}L_{2011}+M_{2006}M_{2011}+G_{2006}G_{2011}]. \quad (13)$$

The stitched model has 1,800 parameters, which is small compared to the 10,077,696 cells of the contingency table that would usually be needed (inaccessible super-table). To calculate the 95% confidence intervals of the prevalence ratios we used a stratified (Pearson) residual bootstrap. In each iteration a new data set was randomly generated from sampling the residuals, and the model estimated using the stitching method. Some confidence intervals were unstable, and were omitted from the table. Bootstrapping techniques for stitched log-linear models with variable cell counts will be developed further in future research.

The mean prevalence ratios of the stitched model (Table 6) are very similar to the best geography model (Table 5) with baseline geography transition remote-to-remote. However, the 95% confidence intervals (that were stable) are generally wider than those from the best geography model. The unstable confidence intervals indicate little information is actually available for that estimate. The difference arises since the confidence intervals in the best geography model are calculated from likelihood profiling

*Table 5. Selected estimated prevalence ratio comparisons and cell counts from the BM6-Geography model. *Indicates baseline of comparison.*

	Prevalence ratios (95% CI) n = cell count					
	Rent to Own					
	21 to 30		31 to 40		41 to 60	
	Male	Female	Male	Female	Male	Female
City → City	87.48 (24.88, 306.00) $n = 4886$	108.54 (30.32, 387.08) $n = 5545$	68.31 (14.24, 320.29) $n = 3988$	65.70 (13.76, 306.24) $n = 4114$	44.51 (10.30, 188.90) $n = 2772$	47.64 (11.27, 198.60) $n = 2608$
City → Region	3.95 (1.14, 13.68) $n = 225$	4.78 (1.37, 16.59) $n = 240$	2.70 (0.43, 16.49) $n = 156$	2.21 (0.34, 13.80) $n = 140$	2.18 (0.50, 9.24) $n = 126$	2.70 (0.62, 11.53) $n = 158$
City → Remote	0.65 (0.46, 0.92) $n = 34$	0.71 (0.50, 1.01) $n = 39$	0.38 (0.26, 0.56) $n = 21$	0.36 (0.24, 0.53) $n = 24$	0.36 (0.25, 0.53) $n = 22$	0.52 (0.35, 0.74) $n = 29$
Region → City	5.50 (1.61, 18.72) $n = 333$	8.25 (2.42, 28.14) $n = 395$	4.73 (1.13, 19.70) $n = 266$	4.43 (1.09, 17.92) $n = 287$	3.35 (0.85, 13.00) $n = 191$	3.44 (0.86, 13.64) $n = 206$
Region → Region	22.16 (6.67, 73.52) $n = 1242$	29.28 (8.96, 95.81) $n = 1491$	19.61 (3.58, 106.16) $n = 1146$	18.72 (3.43, 101.47) $n = 1171$	16.45 (4.29, 62.53) $n = 1026$	18.00 (4.49, 71.55) $n = 984$
Region → Remote	0.75 $n = 41$ (0.54, 1.03)	1.04 $n = 54$ (0.76, 1.43)	0.71 $n = 37$ (0.52, 0.97)	0.98 $n = 66$ (0.73, 1.31)	0.62 $n = 37$ (0.45, 0.86)	0.57 $n = 33$ (0.41, 0.80)
Remote → City	0.33 (0.21, 0.50) $n = 20$	0.32 (0.21, 0.49) $n = 15$	0.16 (0.09, 0.26) $n = 11$	0.14 (0.08, 0.24) $n = 7$	0.19 (0.11, 0.30) $n = 14$	0.22 (0.13, 0.36) $n = 10$
Remote → Region	0.33 (0.22, 0.50) $n = 18$	0.42 (0.28, 0.63) $n = 22$	0.11 (0.05, 0.20) $n = 4$	0.06 (0.03, 0.11) $n = 6$	0.22 (0.13, 0.35) $n = 16$	0.17 (0.10, 0.28) $n = 7$
Remote → Remote	1* $n = 60$	1* $n = 47$	1* $n = 64$	1* $n = 57$	1* $n = 64$	1* $n = 53$

Table 6. Selected estimated prevalence ratio comparisons from combined log-linear model (stitched) related to geography component of the model.

Prevalence ratios (95% CI)						
	Rent to own					
	21 to 30		31 to 40		41 to 60	
	Male	Female	Male	Female	Male	Female
City → City	81.43†	117.98 (0.00, 2339.23)	62.31†	72.18 (22.76, 671.63)	43.31†	49.21†
City → Region	3.75†	5.11 (0.92, 55.05)	2.44†	2.46 (0.65, 23.00)	1.97†	2.98†
City → Remote	0.57†	0.83 (0.00, 10.20)	0.33†	0.42 (0.13, 3.75)	0.34†	0.55†
Region → City	5.55†	8.4 (2.24, 94.31)	4.16†	5.04 (1.87,44.90)	2.98†	3.89†
Region → Region	20.7†	31.72 (6.11,383.34)	17.91†	20.54 (1.38, 268.53)	16.03†	18.57†
Region → Remote	0.68†	1.15 (0.11, 13.24)	0.58†	1.16 (0.13, 12.02)	0.58†	0.62†
Remote → City	0.33†	0.32 (0.08, 3.00)	0.17†	0.12 (0.01, 1.07)	0.22†	0.19†
Remote → Region	0.30†	0.47 (0.15, 4.00)	0.06†	0.11 (0.00, 0.94)	0.25†	0.13†
Remote → Remote	1*	1*	1*	1*	1*	1*

*Indicates baseline of comparison.
†Indicates that confidence intervals were unstable.

rather than bootstrapping, and come from a model which is more likely to overfit the data given that the relative number of parameters versus observations is high (162 vs 216), even after AIC model selection. Even if bootstrapping were undertaken on the best geography model, the low degrees of freedom will dictate smaller residual sizes and hence smaller confidence intervals. As such, the stitched model with bootstrapping is a more robust analysis given the data available to us. It shows that there is actually inconclusive evidence for many of the categories we made inference about in Subsection 4.3.

There are several inferences from Table 6 that can be made. Females in the 31–40 age bracket have high odds of becoming homeowners, especially those staying in the city, staying in regional areas, or moving from regional to city areas. Females in the 21–30 age group that are likely to become homeowners are those staying in regional areas and moving from regional to city areas. The mean prevalence ratio for females aged 21–30 staying in the city was the highest estimate in Table 6 but had a large confidence interval.

5. Conclusion

Decomposition and combination of large log-linear models has been used in work by Dahinden et al. (2010). We adapt this approach to the scenario where contingency table output is restricted from table builders to a set of overlapping marginal tables. We also discuss how to compare these separate marginal models appropriately, but find that in our example the stitched model is more robust.

Table 6 is one of many outputs that can be derived from the stitched model in our example using housing tenure transitions. There are the other variables, and other housing transition categories to consider. Different baselines can also be chosen, which emphasises certain comparisons. We have presented Table 6 since it best relates to our research question about how Australians move from renting to owning. Defining a research question is very important in this analysis (as always) because it determined which tables to request from TableBuilder, which results to extract from our stitched model, and how to display these results.

This article contributes to the toolbox of applied statisticians and researchers to make better use of tabular data where access is subject to query restrictions. It is particularly useful to national statistical agencies (and users of their data sets) who are required to preserve privacy and implement disclosure controls. Future research may address whether similar methods can be used for over- or under-dispersed count data.

6. References

ABS. 2012. TableBuilder user manual. Technical report, Australia Bureau of Statistics, Canberra, ACT (cat.no 2065.0). Available at: http://www.abs.gov.au/tablebuilder (accessed October 2016).

ABS. 2013. *Australian Census Longitudinal Dataset: Methodology and quality assessment – 2080.5 – 2006-11*. Technical report, Australia Bureau of Statistics, Canberra, ACT. Available at: https://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/2080.5Main+Features12006-2016 (accessed October 2016).

Agresti, A. 1981. "Measures of nominal-ordinal association." *Journal of the American Statistical Association* 76(375): 524–529. DOI: https://doi.org/10.1080/01621459.1981.10477679.

Agresti, A. 2002. *Categorical Data Analysis*. Springer, second edition.

Akaike, H. 1974. "A new look at the statistical model identification." *IEEE Transactions on Automatic Control* 19(6): 716–723. DOI: https://doi.org/10.1109/TAC.1974.1100705.

Allison, P.D. 1980. "Analyzing collapsed contingency tables without actually collapsing." *American Sociological Review* 45(1): 123–130. DOI: https://doi.org/10.2307/2095247.

Bergsma, W., M.A. Croon, and J.A. Hagenaars. 2009. *Marginal models: For dependent, clustered, and longitudinal categorical data*. Springer Science & Business Media.

Bergsma, W.P. and T. Rudas. 2002. "Marginal models for categorical data." *Annals of Statistics* 30(1): 140–159. DOI: https://doi.org/10.1214/aos/1015362188.

Birch, M. 1963. "Maximum likelihood in three-way contingency tables." *Journal of the Royal Statistical Society. Series B (Methodological)* 25: 220–233. Available at: https://www.jstor.org/stable/2984562 (accessed November 2017).

Bishop, Y., S. Fienberg, and P. Holland. 1975. *Discrete multivariate analysis: Theory and practice*. The MIT Press, Cambridge, Massachusetts.

Cameron, A. and P. Trivedi. 1998. *Regression analysis of count data*. Cambridge University Press.

Chipperfield, J., D. Gow, and B. Loong. 2016. "The Australian Bureau of Statistics and releasing frequency tables via a remote server." *Statistical Journal of the IAOS* 32(1): 53–64. DOI: https://doi.org/10.3233/SJI-160969.

Chipperfield, J., J. Brown, and N. Watson. 2017. "The Australian Census Longitudinal Dataset: using record linkage to create a longitudinal sample from a series of cross-sections." *Australian and New Zealand Journal of Statistics* 59(1): 1–16. DOI: https://doi.org/10.1111/anzs.12177.

Dahinden, C., M. Kalisch, and P. Bühlmann. 2010. "Decomposition and model selection for large contingency tables." *Biometrical Journal* 52(2): 233–252. DOI: https://doi.org/10.1002/bimj.200900083.

Darroch, J.N., S.L. Lauritzen, and T.P. Speed. 1980. "Markov fields and log-linear interaction models for contingency tables." *The Annals of Statistics* 8(3): 522–539. DOI: https://doi.org/10.1214/aos/1176345006.

Domingo-Ferrer, J. and J. Mateo-Sanz. 1999. "Resampling for statistical confidentiality in contingency tables." *Computers & Mathematics with Applications* 38(11–12): 13–32. DOI: https://doi.org/10.1016/S0898-1221(99)00281-3.

Duncan, G., M. Elliot, and J.-J. Salazar-González. 2011. *Statistical Confidentiality: Principles and Practice*. Statistics for Social and Behavioral Sciences. Springer, New York, NY, second edition.

Frydenberg, M. 1990. "Marginalization and collapsibility in graphical interaction models." *The Annals of Statistics* 8(2): 790–805. DOI: https://doi.org/10.1214/aos/1176347626.

Frydenberg, M. and S.L. Lauritzen. 1989. Decomposition of maximum likelihood in mixed graphical interaction models. *Biometrika* 76(3): 539–555. DOI: https://doi.org/10.2307/2336119.

Jones, E. and V. Didelez. 2017. "Thinning a triangulation of a Bayesian network or undirected graph to create a minimal triangulation." *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 25(3): 349–366. DOI: https://doi.org/10.1142/S0218488517500143.

Lang, J.B. 1996. "On the comparison of multinomial and Poisson log-linear models." *Journal of the Royal Statistical Society. Series B (Methodological)* 58(1): 253–266. Available at: https://www.jstor.org/stable/2346177 (accessed October 2017).

Lauritzen, S.L. 1996. *Graphical models*, volume 17. Clarendon Press.

Lee, J.Y., J.J. Brown, and L.M. Ryan. 2017. "Sufficiency revisited: Rethinking statistical algorithms in the big data era." *The American Statistician* 71(3): 202–208. DOI: https://doi.org/10.1080/00031305.2016.1255659.

Leimer, H.-G. 1993. "Optimal decomposition by clique separators." *Discrete mathematics* 113(1–3): 99–123. DOI: https://doi.org/10.1016/0012-365X(93)90510-Z.

Nelder, J. and R. Wedderburn. 1972. "Generalized linear models." *Journal of the Royal Statistical Society. Series A (General)* 135(3): 370–384. DOI: https://doi.org/10.2307/2344614.

Olesen, K.G. and A.L. Madsen. 2002. "Maximal prime subgraph decomposition of Bayesian networks." *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 32(1): 21–31. DOI: https://doi.org/10.1109/3477.979956.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 2016. Available at: https://www.R-project.org/ (accessed November 2018).

Rose, D.J., R.E. Tarjan, and G.S. Lueker. 1976. "Algorithmic aspects of vertex elimination on graphs." *SIAM Journal on computing* 5(2): 266–283. DOI: https://doi.org/10.1137/0205021.

Spallek, M., M. Haynes, and A. Jones. 2014. "Holistic housing pathways for Australian families through the childbearing years." *Longitudinal and Life Course Studies* 5(2): 205–226. DOI: https://doi.org/10.14301/llcs.v5i2.276.

Received June 2019
Revised October 2019
Accepted February 2020

sciendo

Journal of Official Statistics, Vol. 36, No. 2, 2020, pp. 297–314, http://dx.doi.org/10.2478/JOS-2020-0016

Switching Between Different Non-Hierachical Administrative Areas via Simulated Geo-Coordinates: A Case Study for Student Residents in Berlin

Marcus Groß[1], *Ann-Kristin Kreutzmann*[1], *Ulrich Rendtel*[1], *Timo Schmid*[1], *and Nikos Tzavidis*[2]

The transformation of area aggregates between non-hierarchical area systems (administrative areas) is a standard problem in official statistics. For this problem, we present a proposal which is based on kernel density estimates. The approach applies a modification of a stochastic expectation maximization algorithm, which was proposed in the literature for the transformation of totals on rectangular areas to kernel density estimates. As a by-product of the routine, one obtains simulated geo-coordinates for each unit. With the help of these geo-coordinates, it is possible to calculate case numbers for any area system of interest. The proposed method is evaluated in a design-based simulation based on a close-to-reality, simulated data set with known exact geo-coordinates. In the empirical part, the method is applied to student resident figures from Berlin, Germany. These are known only at the level of ZIP codes, but they are needed for smaller administrative planning districts. Results for (a) student concentration areas and (b) temporal changes in the student residential areas between 2005 and 2015 are presented and discussed.

Key words: Choropleth maps; kernel density estimation; statistical reporting; sub-regional estimation; urban development.

1. Introduction

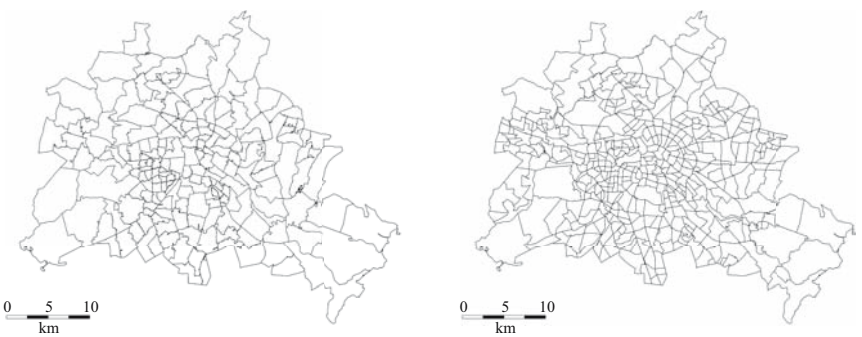
Maps are increasingly used for the dissemination of official statistics. Mostly, these consist of areas that display some value of interest in different colors. Thus, maps demonstrate, for instance, where the poor live (U.S. Census Bureau 2017), where people are most exposed to air pollution (Spiekermann and Wegener 2003), and where accessibility to services is low (Langford et al. 2008; Schmid et al. 2017). Therefore, maps are also an illustrative and easily understandable basis for targeting policies.

The commonly used maps are "choropleths" that use a discretization of the value of interest. The areas or zones are defined, for example, by administrative districts at different levels or statistical units as the European nomenclature des unités territoriales statistiques (NUTS), see the Statistical Atlas of the European Statistical Yearbook (Eurostat 2018). In using choropleth maps, it is problematic that the size of an area is not properly taken into account, which may lead to misinterpretations. Alternatively, areas can be defined by a

[1] Freie Universität Berlin, Garystraße 21, 14195 Berlin, Germany. Emails: Marcus.Gross@inwt-statistics.de, Ann-Kristin.Kreutzmann@fu-berlin.de, Ulrich.Rendtel@fu-berlin.de, and Timo.Schmid@fu-berlin.de
[2] University of Southampton, Murray Building 58, Highfield Campus, Southampton, UK. Email: N.Tzavidis@soton.ac.uk

rectangular grid of a certain size, say 1 km^2. These maps are often referred to as grid maps, see for an example the German Census atlas (Statistische Ämter des Bundes und der Länder 2015). Gallego et al. (2011) discuss several approaches that can be used to downscale area data to fine-scale raster grids in order to receive maps with a higher resolution and thus precision. Grid or raster data is commonly used in urban planning and simulations (Schürmann et al. 2002; Lautso et al. 2004; Patterson et al. 2011). Geo-coded data enables to create a different type of map that is independent of area definitions. These maps are based on a two dimensional kernel density of the variable of interest. They display each level of the estimated density by a different color. In contrast to choropleths, the color scheme, often ranging from light for low values to dark for high values of the density, is continuous. An example is the Service Map of Helsinki (OpenStreetMap Foundation 2019), where the user can combine different background maps with kernel density estimates of demographic subpopulations, like age groups and ethnic minorities.

In addition to the described downscaling or disaggregation of data to subsets of administrative units, switching between different area definitions/systems is often a challenge in official statistics. Performing statistical inference on an area level without available data while having data for another related area level is also known as spatial change of support (COS) (see e.g., Bradley et al. 2016). This occurs when there are different local planning areas in use, for example, fire brigade districts, schooling districts, hospital districts that are different from the standard administrative areas. For certain large-scale planning projects, such as an airport, the number of inhabitants in the upcoming noisefield of airplanes is of interest. European data in the INSPIRE Knowledge Base (Infrastructure for spatial information in Europe) are often reported on squares of different length. Here it may be necessary to adapt the European units to the local units (European Commission 2019). In the application of this work, the number of student residents in administrative areas of Berlin, "Lebensweltlich orientierte Räume" (LOR), is required by the Berlin Senate Department for Urban Planning and Environment for planning purposes. LORs are the smallest urban planning units for Berlin and have an average area size of around 1.99 km^2. However, the university enrollment registers only provide student totals at the level of ZIP codes with an average area size of around 4.62 km^2. Figure 1 shows the 447 LORs, as well as the 193 ZIP-code areas of Berlin. A careful inspection of the areas reveals many cross-cuttings of the area borders. Figure 2



(a) (b)

Fig. 1. ZIP-code areas of Berlin (a) and administrative planning areas (LORs) (b).

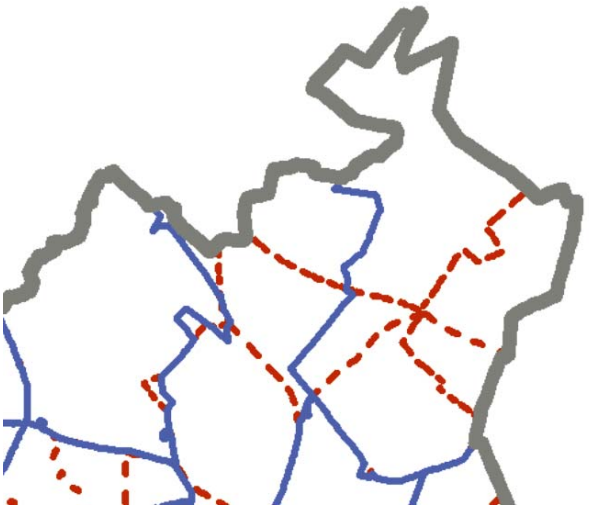


Fig. 2. The non-hierarchical structure of the ZIP areas (blue straight lines) and the LOR areas (red dotted lines) in the north east edge of Berlin.

demonstrates, in detail, the non-hierarchical structure of the ZIP-code areas (blue lines) and the LOR areas (dotted red lines) in the north east edge of Berlin. In other words, LORs are by no means a lower-level area system than ZIP-code areas.

As these two area systems are non-hierarchical, we are confronted with a problem that is hard to solve at an elementary level. Often this task is advanced by ad hoc methods, based on a proportional allocation of totals depending on which part of the ZIP-code area belongs to the respective administrative planning area (LOR). Such an approach is tedious and relies on an unrealistic assumption, namely, that the units are uniformly distributed across the ZIP-code area. Instead, Mugglin and Carlin (1998) and Mugglin et al. (1999) propose a hierarchical Bayesian approach for the spatial change of support. Bradley et al. (2016) extend the approach to data with sampling variability and enable the spatial and temporal change of support. These model-based approaches require covariate information on the area of interest and rely on distributional assumptions. Within the field of small area estimation, Trevisani and Gelfand (2013) extend hierarchical Bayesian models to soften the requirements for the covariate information by allowing the use of covariates of areas non-nested within the small areas of interest. In this work, we suggest a non-parametric alternative in the form of a kernel density estimate (KDE) that tackles the problem of transferring count numbers from one area system to another without covariate information. As the density function is independent of administrative areas, it is possible to compute count numbers for any area definition/system from the density.

In our case, we do not have the exact geo-coordinates at hand but only totals for areas that are not related to the areas of interest. Therefore, we present an approach in which geo-coordinates are simulated from area-specific aggregates. The method proposed in this work is similar to the approach of Groß et al. (2017), where it is used to counteract the rounding of geo-coordinates due to confidentiality reasons. In their analysis, kernel densities are generated to detect concentration areas of migrants and elderly persons in

Berlin. Rendtel and Ruhanen (2018) use the approach with "Open Data" in order to demonstrate local need for child care.

The algorithm of Groß et al. (2017) works for totals on rectangles that are the outcome of the rounding process. However, the approach can be extended to totals of arbitrary shape files. The algorithm is based on two elementary steps: the first step is to draw a sample from a two-dimensional density that gives the simulated geo-coordinates. The sampling is done with respect to the known number of observations in the reference areas, which is achieved by stratified sampling. The second step is a classical estimation step that generates a kernel density estimate from a sample of geo-coded data. These two steps resemble a "stochastic expecation maximization" (SEM) algorithm (Celeux et al. 1996).

The article is organized as follows. In Section 2, the proposed algorithm and its statistical foundation is described in more detail. Section 3 evaluates the quality of the conversion to different areas via a design-based simulation study. The proposed method is applied to the Berlin student residents data set in Section 4. Furthermore, the problem of allocating the students of Berlin to administrative areas/LORs for the planning of student homes and other student-related infrastructure is discussed. Besides the estimation of the total number of students in administrative areas/LORs, the kernel densities offer alternative methods to display regions with a dense student population and their development over time. The method is compared with the classical approaches via choropleths. Section 5 concludes and provides further research ideas.

2. Method

Multivariate kernel density estimation is a non-parametric approach to estimate the joint probability distribution of two or more continuous variables. Let $\boldsymbol{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ denote the exact geo-coordinates, such as longitude and latitude, of observations $i = 1,.., n$ with $\boldsymbol{X}_i = (X_{i1}, X_{i2})$. To estimate the density $f(x)$ at point x, a multivariate kernel density estimator is employed, which is given by:

$$\hat{f}_H(x) = \frac{1}{n|\boldsymbol{H}|^{\frac{1}{2}}} \sum_{i=1}^{n} K\left(\boldsymbol{H}^{-\frac{1}{2}}(x - \boldsymbol{X}_i)\right), \tag{1}$$

where $K(\cdot)$ denotes a multivariate kernel function and $|\boldsymbol{H}|$ denotes the determinant of a symmetric positive definite bandwidth matrix $\boldsymbol{H}$. A popular choice for $K(\cdot)$ is the multivariate Gaussian kernel. The choice of $\boldsymbol{H}$ is highly important for the performance of the kernel density estimator. In principle, all bandwidth selection strategies try to minimize the mean integrated squared error (MISE) which is $E\int\left(\hat{f}_H(x) - f(x)\right)^2 dx$, where $f(x)$ is the true density and $\hat{f}_H(x)$ is the kernel estimate using the bandwidth matrix $\boldsymbol{H}$. For high case numbers, the asymptotic MISE (AMISE) offers some simplification by omitting some terms of lower order. The essential part depends on the mixed derivatives of the underlying density $\int f^{(m)}(x) f(x) dx$. Wand and Jones (1994) suggest some simple but efficient approximations of the empirical substitute $1/n\sum_{i=1}^{n} \hat{f}_H^{(m)}(x_i)$. They discuss the choice of the bandwidth in the multivariate case by using a plug-in estimator, which is also used in this work.

Instead of the exact geo-coordinates $\boldsymbol{X}$, only aggregated data for certain areas is available in this work. Simply applying a kernel density estimator to, for instance the area centers, leads to strongly biased estimates for rectangular shapes as shown in Groß et al. (2017). Therefore, a special treatment is needed. Following Groß et al. (2017), we can interpret the available data on area level, denoted by $\boldsymbol{W} = \{\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n\}$, as a coarse measurement of the exact geo-coordinates of individual i. As the measurement process is known, we are able to formulate a measurement model $\pi(\boldsymbol{W}|\boldsymbol{X})$ for $\boldsymbol{W}$. It can be written as a simple product of Dirac distributions, $\pi(\boldsymbol{W}|\boldsymbol{X}) = \prod_{i=1}^{n} \pi(\boldsymbol{W}_i|\boldsymbol{X}_i)$, with

$$\pi(\boldsymbol{W}_i|\boldsymbol{X}_i) = \begin{cases} 1 & \text{for } \boldsymbol{X}_i \in area(\boldsymbol{W}_i), \\ 0 & \text{else}, \end{cases} \tag{2}$$

where $area(\boldsymbol{W}_i)$ stands for the set of geo-coordinates that belong to the area where $\boldsymbol{W}_i$ lies in.

From $\pi(\boldsymbol{X}_i|\boldsymbol{W}_i)$, we can draw pseudo samples of the $\boldsymbol{X}_i$ to estimate the density f by using the Bayes theorem:

$$\pi(\boldsymbol{X}_i|\boldsymbol{W}_i) \propto \pi\left(\boldsymbol{W}_i|\boldsymbol{X}_i\right)\pi(\boldsymbol{X}_i). \tag{3}$$

Thus, the exact geo-coordinates, $\boldsymbol{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$, are distributed according to the kernel density estimate restricted to the area where the observation $\boldsymbol{W}_i$ comes from. In an iterative procedure, the $\boldsymbol{X}_i$ are sampled from $\pi(\boldsymbol{X}_i|\boldsymbol{W}_i)$ followed by the estimation of $\pi(\boldsymbol{X}_i)$, respectively $f(x)$, by employing a multivariate kernel density estimator on the $\boldsymbol{X}_i$.

In particular, a SEM algorithm (Celeux et al. 1996) is utilized. This is a modification of the original EM-algorithm. The basic setting of the EM-algorithm refers to a situation where a part of the observations is missing. Thus, one has to maximize the marginal loglikelihood for the observed part of the data. As this can be quite complicated, one regards the expected value of the loglikelihood of the complete data where the expectation is done with respect to the current estimate of the parameter estimate. In many instances, this expected value of the loglikelihood of the complete data can be maximized by standard routines and leads to an update of the parameter estimate. With the SEM algorithm, the expected value is replaced by one realization of the unobserved part of the sample under the current parameter estimate. Again, the likelihood for this completed pseudo-sample is maximized and a new update of the parameter estimate is achieved. The generation of the pseudo-sample step brings a stochastic element into the algorithm, giving a more realistic distribution of the missing observations. In our application, the missing data are the exact geo-coordinates and the maximization of a likelihood is replaced by the kernel density estimation (a generalised SEM, Groß et al. 2017). In contrast to SEM, a simple EM algorithm would clearly not be helpful in this application, as all observations within an area would fall on the same location and thus not prevent a bias of the resulting kernel density estimate.

The algorithm starts with all the points concentrated at the center of the area. Starting from these artificial geo-coordinates, a kernel estimate $\hat{f}_{(0)}(x)$ is generated. Two iterative computation steps are performed as follows:

Step 1 (the 'S'-step in SEM): "Pseudo-samples" of the exact geo-coordinates, the $\boldsymbol{X}_i$, are drawn by sampling from the conditional distribution $\pi(\boldsymbol{X}_i|\boldsymbol{W}_i)$. This conditional

distribution is equal to the current density estimate restricted to the area where $\boldsymbol{W}_i$ belongs

Step 2 (the 'M'-step in SEM): The bivariate kernel density $\hat{f}_{(n+1)}(x)$ is estimated using the drawn pseudo-sample.

By $B + N$ iterations of Step 1 and Step 2, a sequence of kernel density estimates is generated. The final density estimate is computed by averaging the estimates of $\hat{f}_{(n)}(x)$ over the N samples after discarding the first B burn-in samples. The number of burn-in samples to achieve convergence and the number of samples N for a desired accuracy may depend on the application. As for MCMC-methods, no general recommendations can be given. However, for similar applications as presented here, we found that $B = 5$ and $N = 100$ was generally sufficient. More details on the kernel density estimation method and the exact implementation of the algorithm can be found in Groß et al. (2017). The only detail that is changed is to draw the pseudo-samples from the corresponding shape rather than from a rectangle, that means in the 'S'- step truncating the density to the area where observation $\boldsymbol{W}_i$ lies. This is more computationally intense, especially for complex-formed shapes, because we have to check whether there is a potential pseudo-sample inside the shape. However, this is of little importance with modern computers as long as the areas do not have a very complex shape, for example non-convex shapes cut into separate parts.

In our application, the problem arises from the fact that large areas of a town may consist of unsettled areas, like parks, lakes or industrial areas. These areas should be excluded from the generation of the geo-coordinates. If this information is available, it may considerably improve the estimation of the case numbers in the new area system. In principle, this is no problem for the SEM algorithm. One simply has to exempt the unsettled areas from the sampling of the geo-coordinates. However, in this case the boundary problem of the kernel density estimation comes into play, as the kernel function may not respect the boundary of the settled regions. One approach, the "cut- and normalize-method" (Gasser and Müller 1979), to overcome this problem is to restrict the kernel function to settled areas and to compute a new normalizing factor that makes the kernel function on the reduced area a density. Such a factor has to be computed for every spot where the kernel function is evaluated. This costs computational time, but it is not a real obstacle as it is implemented in existing software (Groß 2018).

After computing a non-parametric density estimate with this algorithm, the question arises how to allocate the number of observations to each shape in the new target area system. One possibility would be to numerically integrate over the non-parametric density and multiply the result by the number of total observations. However, it is likely that the result would not be consistent with the original data, that is, the number of observations belonging to a shape of the first area level would be different from the starting values. To preserve the original data structure, we chose to count the pseudo-samples falling in each shape of the target area system for each iteration. This also avoids numerical integration. These area counts will be averaged over all iterations.

The existence of N replications of an estimate makes it possible to calculate a confidence interval for the population value. In our simulation study below, we computed an interval that is given by the 5% and the 95% quantile of the N replications. This is not an exact confidence interval as it ignores the sampling from the density. But in our

application, sampling and its induced variance is not an issue as the starting values, that is, the numbers at ZIP-level, are population values with no variance. Nevertheless, the distribution over N replications reflects exactly the uncertainty of the knowledge of the case numbers in the new area system.

The algorithm is implemented in the R package *Kernelheaping* (Groß 2018) as function *dshapebivr*, which requires a data matrix with aggregated observation numbers for each area and a ∗.shp shapefile including the geometric data as input. The function *toOtherShape* in this package performs the operation to preserve the original data structure given the output of the function *dshapebivr* and an additional shapefile for the new administrative area system.

3. Simulation

In order to check the precision of the proposed method, we generated close-to-reality populations in a simulation. As a reminder, the areas of interest are the 447 LORs of Berlin, while the information of student totals is only given at the 193 ZIP-code areas. The cross-cutting of these area systems shown in Figure 3 confirms that the area systems are non-heirarchical.

We then randomly selected 15 mid-points to avoid a simple cluster in the center of the town. At each mid-point, 2,000 observations were generated from bivariate normal density with a variance of 3×10^6 (with covariance equal to 0). Then the points that were allocated to uninhabited areas were removed. Afterwards, we used two versions of the SEM algorithm. In the first version, we ignored the information about which areas are unsettled (SEM), while in the second version, we used the boundary correction (SEM-Boundary)

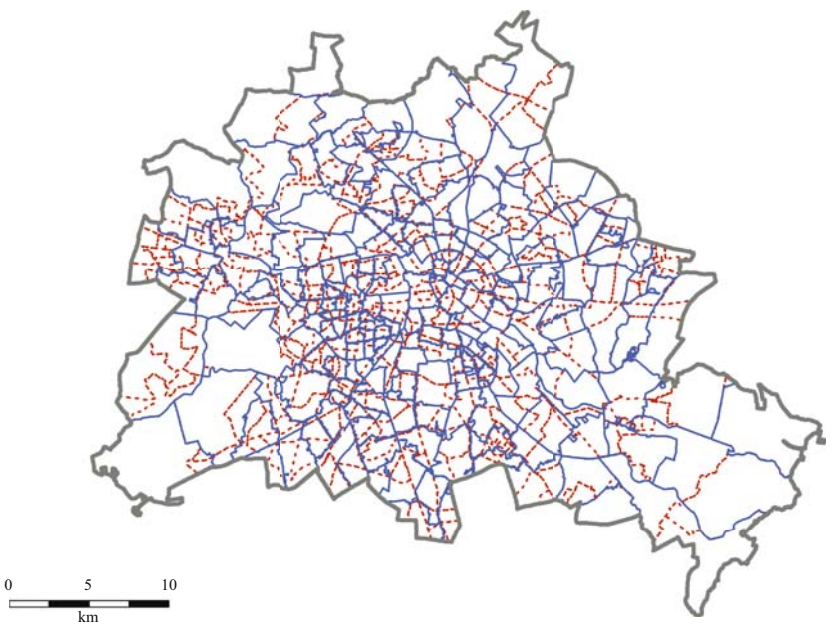


Fig. 3. Cross-cutting of ZIP-code area (blue, straight lines) and LOR area (red, dashed lines) borders in Berlin.

that keeps the density estimate within the settled areas. In order to evaluate the routine against the standard GIS procedure, we used a uniform density within the areas. Here, we used two versions as well. The first version ignores the unsettled areas (UNIFORM), while the second version respects the unsettled areas (UNIFORM-Boundary). The uniform allocation of the observations within the ZIP-code areas avoids the tedious computation of the cross-cutted areas that would be necessary in the GIS-approach, but is approximately equivalent. This procedure was repeated $R = 100$ times, however the preselected 15 mid-points were kept fixed. In order to compute 90-percent confidence regions, we selected the number of replications as $N = 100$. The burn-in phase was taken as $B = 5$.

Figure 4 displays one artificial allocation of geo-coordinates together with the LOR borders (left) and with the unsettled areas in green (forests and parks), blue (water) and grey (industrial and other).

In a next step, the number of observations falling in each area is counted at LOR-area level (treated as true values) and at the ZIP-code area level. The ZIP-code area level counts are used to estimate the "true" counts at the LOR-area level. As explained in Section 2, this is done by counting the number of the generated pseudo-samples falling in each LOR. There is no extra computational effort: during the generation of a new density, it can be checked in which of the LORs the new coordinates fall. Hence, every round of the SEM algorithm produces an estimate of the expected number of points falling into a LOR. Thus, it is only necessary to average these figures over the N Monte-Carlo replications.

Table 1 compares the performance of the four procedures with respect to the root mean squared error (RMSE) of the estimated LOR totals over the R replications, defined as

$$RMSE = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\widehat{LOR}_r - LOR_r\right)^2}$$

where $\widehat{LOR}$ denotes the estimated and LOR the true LOR total. The RMSE is computed for every area and the distribution of these area-specific RMSE values is then analyzed over areas. We see that the information on settled areas is helpful in reducing the RMSE

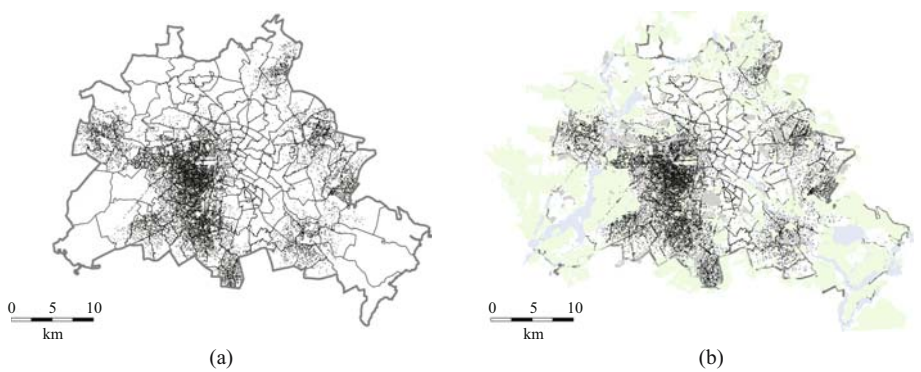


Fig. 4. Simulated geo-coordinates (a) and including their restriction to settled areas (b). Result of one out of 100 simulation runs.

Table 1. RMSE and coverage of the estimated LOR totals over the R = 100 simulation runs.

Method	Average RMSE	95% quantile RMSE	99% quantile RMSE	Max RMSE	Coverage of 90% quantile
SEM	5.63	9.91	24.85	45.4	83.54
UNIFORM	7.33	11.88	34.03	63.6	N/A
SEM-Boundary	4.36	7.45	15.93	38.70	90.54
UNIFORM-Boundary	5.04	9.19	16.24	47.05	N/A

for both algorithms. However, the average RMSE of the SEM algorithm is always lower than the average RMSE with the UNIFORM procedure. With no information on settled areas, the differences are more pronounced: here the SEM algorithm amounts to only 76.8% of the RMSE with the uniform distribution. With information on settled areas, the reduction drops to 86.6%. Similar figures are obtained for the upper quantiles of the RMSE.

The last column displays the coverage of the 90% quantile interval based on the replications of the SEM algorithm. For each area count there is such an interval. The area-specific coverage rates are then averaged over all areas. For our simulations, the average coverage of this interval is close to its nominal value. Thus, the variation of the N replications of the SEM algorithm may be used to construct a confidence interval for the area counts.

In order to demonstrate the use of the *Kernelheaping* package, we present a minimal working example in the Github repository Kernelheaping MinimalWorkingExample.

4. Application

The city of Berlin is a growing town. In the past five years, Berlin has gained around 220,000 people in total, see Senator für Stadtentwicklung und Umwelt (2016). A large proportion of this increase is due to the population gains in the age group of 20 to 30 years old, which contains many students. With the increasing number of students, questions for urban development planning arise. Where do students live and how do they get to their universities? What type of housing do students need? Students, as well as other social groups, have special requirements and behavioral patterns with regard to the local infrastructures.

To answer the above questions, it is helpful to have accurate and reliable information of the residential locations of students in Berlin. This information can improve the planning of projects that students benefit from and, consequently, these can be implemented more targeted. Therefore, the Senate Department for Urban Development and Environment aimed to analyze where students who are enrolled at Berlin universities are located in the metropolitan region of Berlin-Brandenburg and how they relate to the counts of LORs and Brandenburg municipalities. Before, there was no data available about student locations at small-scale residential areas. The LORs are the smallest urban planning units in Berlin. One possible data source about student residences are the enrollment offices of the Berlin universities. However, for privacy concerns, these figures are available only at the level of ZIP-code coordinates.

4.1. The Data

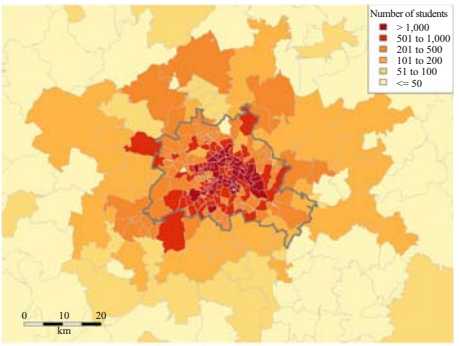
The number of students at ZIP-code area level in the years 2005, 2010 and 2015 could be established for the three – by far largest – universities of Berlin: Freie Universität Berlin (FU), Humboldt-Universität zu Berlin (HU) and Technische Universität Berlin (TU). The same applies for the rather small Alice Salomon Hochschule Berlin. Only for the year 2015, we were provided with numbers from Beuth Hochschule für Technik Berlin, the Hochschule für Wirtschaft und Recht Berlin (HWR) and the Hochschule für Technik und Wirtschaft Berlin (HTW). All numbers refer to the beginning of the winter term ('Wintersemester', abbr. WS), except for the data from FU and HU in 2015, which refer to the summer term ('Sommersemester', abbr. SoSe), where student numbers are typically lower. Thus, we applied a correction for the HU and the FU in 2015 and multiplied the numbers of these two universities by the ratio of winter term to summer term (e.g., FU: $36{,}674 = 33{,}173 \cdot 1.106$). Table 2 gives an overview of the total number of students in each year for every college and university, as well as the total number of students in Berlin. The data is provided by the Statistical Office for Berlin-Brandenburg. Figure 9 (see Appendix 6) visualizes the locations and size of the colleges and universities in Berlin. Furthermore, we have information on all dormitories in Berlin and the number of students living there for every considered year. As our information on ZIP-code totals does not cover all educational institutes with students in Berlin, our totals only sum up to 80% of the total number of students. With respect to the total number of students in Berlin, there is precise information from official statistical sources (Amt für Statistik Berlin-Brandenburg 2018). In order to cover the rest of the students from other institutes, we used some calibrations for the ZIP-code totals. As this calibration is not relevant for the method displayed here, we deferred the details of our calibrations to the appendix (see Appendix, Section 6). The students living in dormitories are not used for the kernel density estimates and are added afterwards to the final estimates at LOR-level to produce more accurate estimates, as their location is already known.

4.2. Results for the Location of Students in Different Map Representations

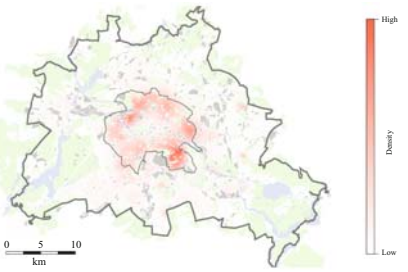
The maps in Figure 5 visualize the number of students in ZIP-code area (the level for which data is available), the kernel density estimate (transmission tool) computed on the

Table 2. Number of students in 2005, 2010 and 2015 for available colleges.

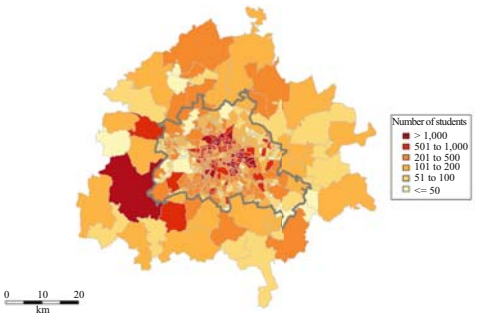
College/University	WS 2005	WS 2010	WS 2015	SoSe 2015
TU Berlin	29,772	29,758	33,933	-
FU Berlin	34,936	33,518	36,674	33,173
HU Berlin	32,428	29,689	34,214	31,098
Beuth	-	-	12,532	-
HTW	-	-	13,355	-
Alice Salomon	1,611	2,512	3,422	-
HWR	-	-	10,009	-
Sum of available colleges	98,697	95,477	144,139	-
Sum of all Berlin colleges	133,024	147,030	175,651	-



(a) Distribution at ZIP-code level



(b) Kernel density estimates



(c) Distribution at LOR level

Fig. 5. The plots show number (density) of Berlin students in 2015.

basis of these counts and the estimated number of students in the LORs of Berlin (the level of interest for urban planning) and its surrounding municipalities in 2015.

All three maps display a joint pattern with a concentration of students in a belt surrounding the center of the town. This belt is characterized by a traditional dense settlement (Senate Department for Urban Development and Housing 2017). It can also be seen that some students commute from neighboring municipalities to Berlin universities. Clearly, this number declines rapidly with the distance from Berlin. However, the graphical impression of the map with ZIP-code areas and LORs is quite different in the southwest of Berlin (the area of Potsdam). In the LOR representation, it looks very much as if there is a cluster that is densely populated with students. However, the ZIP-code area and the KDE representation do not exhibit such a pattern. The southwest "cluster" is simply the result of taking the entire municipality of Potsdam as one LOR.

When it comes to the individual development of the LORs with the highest student counts, it can be noted that they are located in certain districts of Berlin (Wedding, Neukölln, Moabit, Prenzlauer Berg, Friedrichshain and Kreuzberg). Table 3 lists the ten most popular LORs among students in 2015 and their development over time together with the 95% coverage interval for the 2015 values. The values exhibit remarkable changes in their student population from 2005 to 2015. Thus, the necessity of studies aiming to monitor the changes of the student population at a low level of regional aggregation is restated. With the exception of Rixdorf in Neukölln, all areas with a substantial increase of the student population lie in the north-west (Wedding and Moabit) of the central belt. In all the other LORs the student population is quite stable over time.

Table 3. The ten most popular urban planning areas (LOR) in 2015 with students counts for 2005, 2010 and 2015. The 95% coverage interval refers to the year 2015.

Urban planning area	District	2015	Coverage Interv.	2010	2005
Reuter Kiez	Neukölln	2072	(2051, 2094)	2187	1956
Samariterviertel	Friedrichshain	1892	(1833, 1940)	2095	2159
Rixdorf	Neukölln	1856	(1805, 1899)	1469	869
Rehberge	Wedding	1680	(1630, 1725)	1148	773
Traveplatz	Friedrichshain	1637	(1571, 1704)	1226	1354
Emdener Straße	Moabit	1580	(1566, 1591)	1162	942
Soldiner Straße	Wedding	1540	(1521, 1565)	1036	695
Reinickendorfer Straße	Wedding	1440	(1382, 1493)	923	603
Graefe Kiez	Kreuzberg	1409	(1393, 1426)	1350	1513

4.2.1. A Closer Look at the Temporal Development 2005–2015

Since data is available for the years 2005, 2010 and 2015, we can have a closer look at the temporal development of Berlin students residencies.

In general, the proportion of students living in Berlin has slightly but steadily increased from 82.3% in 2005 to 84.4% in 2015. In contrast to that, the percentage of students living in other German regions and foreign countries (mostly Poland) has decreased from 7.1% in 2005 to 5.0% in 2015. For a more detailed overview, Table 4 shows the estimated proportions of students living in Berlin, in the surrounding municipalities, in other municipalities of Brandenburg and outside of Berlin or Brandenburg. Focusing on Berlin and its surroundings, Figure 6 shows the KDE maps for each of the three reference years 2005, 2010 and 2015. From this representation, the structure of the students settlement seems to remain quite stable. However, if we display the highest density regions (HDR) remarkable regional changes can be seen. Note that such a representation is restricted to the KDE approach. Figure 7 compares the HDRs containing 25% and 50% of the students over time. Parts of the northwestern inner belt (Moabit and Wedding), as well as the southern belt (Neukölln) are now included in the 25% region in comparison to 2005. Parts of the eastern belt (southern Prenzlauer Berg and parts of Friedrichshain and Kreuzberg) did drop off from the 25% HDR in the last ten years. Interestingly, it becomes apparent that, in general, the concentration decreased. The 25% highest density region enfolded only 24.64 km^2 in 2005. This area grew to 28.58 km^2 in 2010 and to 33.27 km^2 in 2015. A

Table 4. Distribution of students of Berlin colleges living in Berlin, in the surrounding municipalities, in other municipalities of Brandenburg and out of Berlin/Brandenburg.

	2005	2010	2015
Berlin	109,436 (82.3%)	121,356 (82.5%)	148,231 (84.4%)
Surrounding municipalities	6,713 (5.0%)	7,648 (5.2%)	9,595 (5.5%)
Other municipalities of Brandenburg	7,504 (5.6%)	8,620 (5.9%)	9,059 (5.2%)
Other German regions and foreign countries	9,470 (7.1%)	9,406 (6.4%)	8,766 (5,0%)
Overall	133,024 (100%)	147,030 (100%)	175,651 (100%)

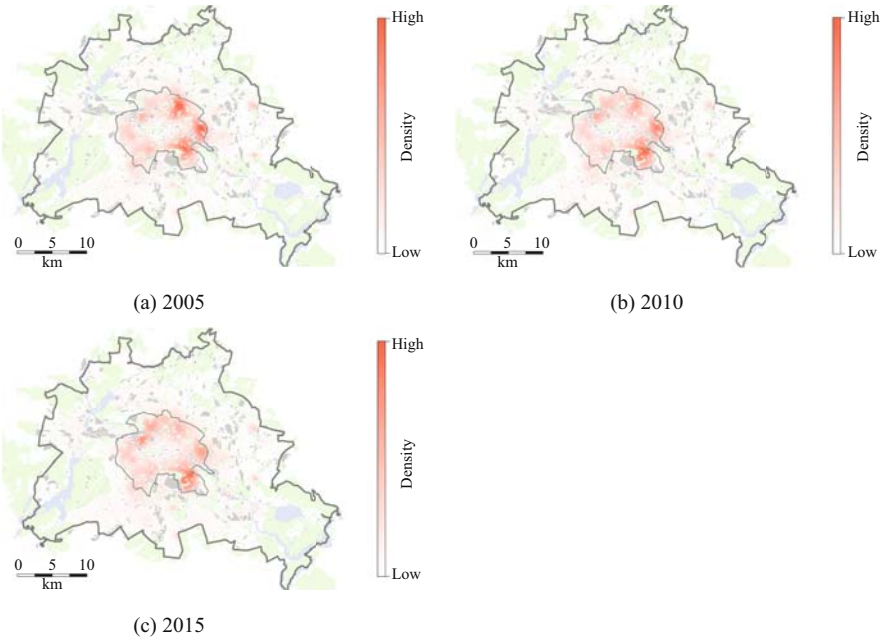


Fig. 6. The plots show the kernel density estimates of Berlin students in 2005 (a), 2010 (b) and 2015 (c).

similar effect is noticeable for the 50% HDR (2005: 76.88 km^2, 2010: 81.45 km^2, and 2015: 92.40 km^2).

Analyzing the absolute differences in the number of students on the level of the urban planning areas reveals further insights. Differences over the whole time period are visualized in Figure 8. A very large increase can be observed here for the locality of Wedding (northwest). The localities Neukölln (south), Lichtenberg (east), Moabit (northwest) and to a lesser extent Adlershof (southeast), Tempelhof (south) or Schöneberg (southwest) have gained students. Strong negative trends are recorded for Prenzlauer Berg (northeast) and the northern part of Mitte (center), which can be attributed to the

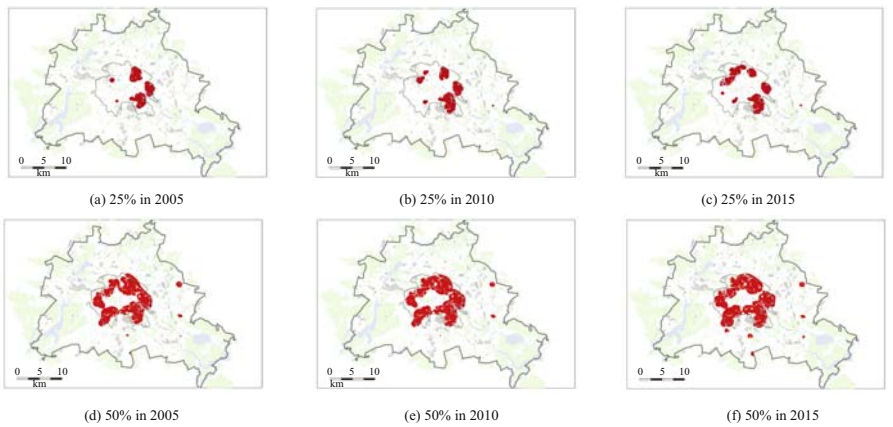


Fig. 7. Regions with highest student density: 25% of students (a, b, c) and 50% of students (d, e, f).

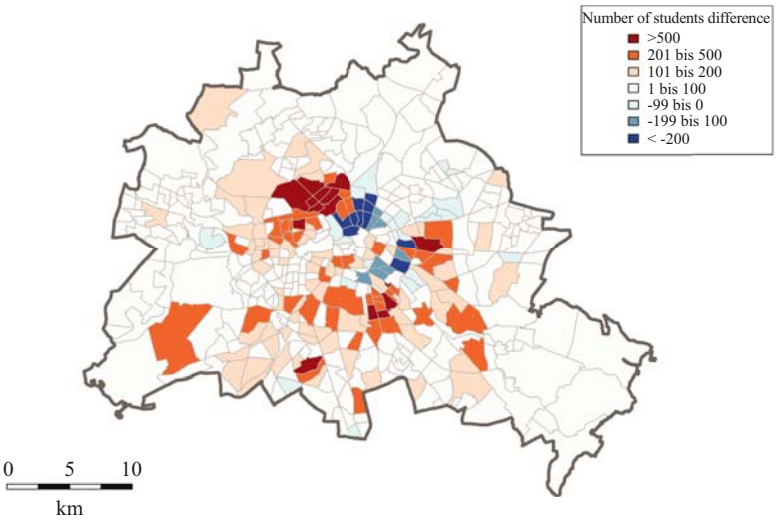


Fig. 8. Differences in student numbers 2015 compared to 2005 on administrative planning area level.

gentrification of these quarters (Schulz 2017; Holm and Schulz 2018). In addition, the eastern parts of Friedrichshain (east) and Kreuzberg (southeast) have lost students in the reference period.

The observations described may be due to the general increase of student numbers by almost 32% in Berlin, see Table 4. But they are also the result of a tightening housing market, which led the students to search for an apartment in other areas where housing is affordable for them. By contrast, Moabit, Wedding and Neukölln are propagated in the discussion on revaluation and displacement processes that can be carried out by pioneers such as students.

5. Conclusion

This work shows that kernel density estimates are a useful tool for the transformation of case numbers between area systems that are not hierarchical. Compared to ad hoc solutions, the proposed method is particularly preferable due to the following reasons. First, our approach is not based on the unrealistic assumption that the characteristic is uniformly distributed within areas. Second, while ad hoc solutions are often carried out manually, the approach in this work is available in the R package *Kernelheaping* and thus, the user can do this task quite automatically. Third, the algorithm is able to deal with uninhabited areas, which is a problem that is often encountered in practice. Fourth, the algorithm delivers coverage intervals for the population values. Finally, the proposed method is superior to the ad-hoc approach with respect to the RMSE.

Furthermore, density estimates, which are used as a transmission tool in this work, have their own merits. They help highlight the highest density regions, which can be used to identify local concentrations in the region of interest.

It should be noted that our algorithm is extremely useful for the construction of maps that are based on Open Data. Because of confidentiality reasons and their easy access, they

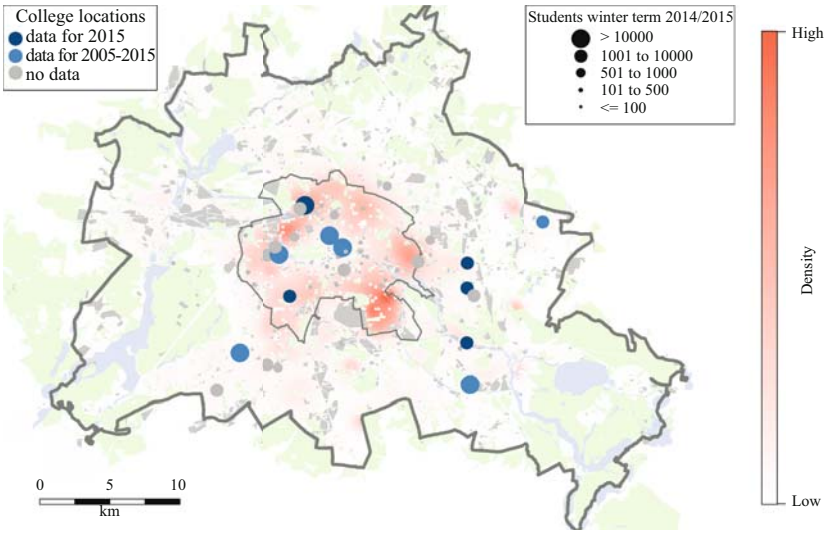


Fig. 9. Locations of colleges and universities of Berlin with number of students including the kernel density estimate of the student distribution in 2015. The border of the 'inner city' is added to the map.

are often provided as local aggregates. For example, the Open Data in Berlin are presented at the level of LORs or at a grid level (Berlin Open Data 2019). In the considered application the disclosure risk of individuals is not increased as the simulated geo-coordinates of individuals of a certain ZIP-code area are all drawn from the same distribution. However, additional information at individual level, such as ethnic affiliation, might help to identify an individual's location more precisely by running the presented algorithm on different sub-groups.

6. Appendix

The vast majority (about 80%) of Berlin's students in 2015 was covered by our sample of colleges and universities. Nevertheless, we would clearly underestimate the number of students in the planning areas due to the missing colleges. Therefore, a calibration is

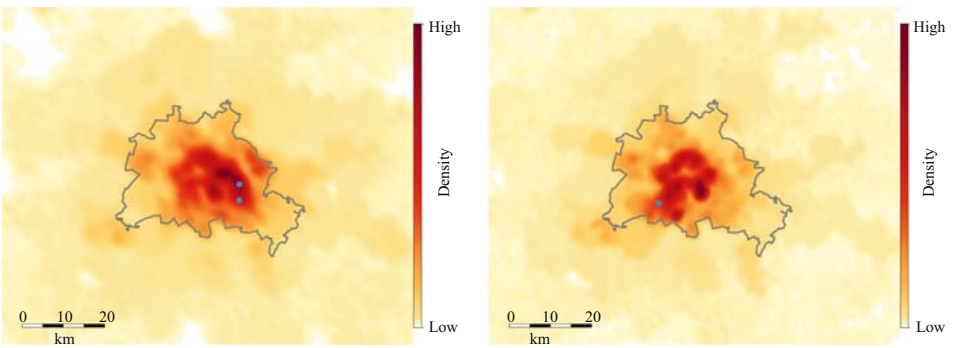


Fig. 10. Kernel density estimates of HTW (a) and FU (b) student distributions with college site locations.

necessary. The Statistical Office for Berlin-Brandenburg provides the total numbers of students enrolled in Berlin, giving us the possibility to simply upscale the total number of students in each ZIP-code area by a factor (e.g., multiplying by 175,651/144,139 = 1.22 for 2015; cf. Table 2). Another issue is the problematic comparison of the years 2005 and 2010 with 2015, as the coverage of colleges and universities is lower in these years. This is especially important as the specific college has a definite influence on the students' living address. We found out that a large proportion of the students live within the inner city borders, but some live near the college as well, as the kernel density estimate for 2015 shows (cf. Figure 9).

For the year 2015, we think that the effect of missing colleges is negligible, as we have information on the most important ones and the remaining ones are rather small and quite similarly distributed. If we would leave out the colleges only available in 2015, we get quite different area aggregates for ZIP-code areas near the missing colleges, for example ZIP code 10318 with only 145 instead of 796 students. Figure 10 excellently shows the kernel density estimates of the HTW and the FU student distributions. To account for the lower number of colleges in 2005 and 2010, we tried to adjust the number of students using the data of 2015. To achieve this, we employed a generalized linear mixed model, GLMM, (McCulloch and Neuhaus 2001) linking the number of students in each ZIP-code area considering all colleges available (Y) with the number considering colleges with data available for 2005 to 2015 (X). With a random intercept for each ZIP code ($zip_i \sim N(0, \tau)$), we fitted a Poisson-glmm with a log-link and the following model formula:

$$Y_i = \exp(\beta_0 + log(X_i + 1)\beta_1 + zip_i)$$

This formula was then used to predict Y for the years 2005 and 2010.

7. References

Amt für Statistik Berlin-Brandenburg. 2018. *Statistiken zu Bildung und Kultur*. Available at: https://www.statistik-berlin-brandenburg.de/BasisZeitreiheGrafik/Zeit-Hochschulen.asp?Ptyp=400&Sageb=21003&creg=BBB&anzwer=7 (accessed February 2019).

Berlin Open Data. 2019. *Datensätze*. Available at: https://daten.berlin.de/datensaetze (accessed February 2019).

Bradley, J.R., C.K. Wikle, and S.H. Holan. 2016. "Bayesian spatial change of support for count-valued survey data with publication to the American Community Survey." *Journal of the American Statistical Association* 111(514): 472–487. DOI: https://doi.org/10.1080/01621459.2015.1117471.

Celeux, G., D. Chauveau, and J. Diebolt. 1996. "Stochastic versions of the EM algorithm: an experimental study in the mixture case." *Journal of Statistical Computation and Simulation* 55(4): 287–314. DOI: https://doi.org/10.1023/B:STCO.0000039481.32211.5a.

European Commission. 2019. *Inspire knowledge base - infrastructure for spatial information in Europe*. Available at: https://inspire.ec.europa.eu/ (accessed May 2019).

Eurostat. 2018. *Statistical Atlas*. Available at: http://ec.europa.eu/eurostat/statistical-atlas/gis/viewer/# (accessed January 2019).

Gallego, F.J., F. Batista, C. Rocha, and S. Mubareka. 2011. "Disaggregating population density of the European Union with CORINE land cover." *International Journal of Geographical Information Science* 25(12): 2051–2069. DOI: https://doi.org/10.1080/13658816.2011.583653.

Gasser, T. and H.-G. Müller. 1979. "Kernel estimation of regression functions." In *Smoothing techniques for curve estimation*: 23–68. Springer.

Groß, M. 2018. *Kernelheaping: Kernel Density Estimation for Heaped and Rounded Data*. R package version 2.2.0. Available at: https://cran.r-project.org/web/packages/Kernelheaping/ (accessed May 2020).

Groß, M., U. Rendtel, T. Schmid, S. Schmon, and N. Tzavidis. 2017. "Estimating the density of ethnic minorities and aged people in Berlin: Multivariate kernel density estimation applied to sensitive geo-referenced administrative data protected via measurement error." *Journal of the Royal Statistical Society: Series A* 180(1): 161–183. DOI: https://doi.org/10.1111/rssa.12179.

Holm, A. and G. Schulz. 2018. GentriMap: "A model for measuring gentrification and displacement." In I. Helbrecht, ed. *Gentrification and Resistance*: 251–277. Wiesbaden: Springer VS.

Langford, M., G. Higgs, J. Radcliffe, and S. White. 2008. "Urban population distribution models and service accessibility estimation." *Computers, Environment and Urban Systems* 32(1): 66–80.

Lautso, K., K. Spiekermann, M. Wegener, I. Sheppard, P. Steadman, A. Martino, R. Domingo, and S. Gayda. 2004. Planning and research of policies land use and transport for increasing urban sustainability. Report, European Commission. Available at: http://www.spiekermann-wegener.de/pro/pdf/PROPOLIS_Final_Report.pdf (accessed February 2019).

McCulloch, C.E. and J.M. Neuhaus. 2001. *Generalized linear mixed models*. Wiley Online Library. DOI: https://doi.org/10.1002/9781118445112.stat07540.

Mugglin, A.S. and B.P. Carlin. 1998. "Hierarchical modeling in Geographic Information Systems: Population interpolation over incompatible zones." *Journal of Agricultural, Biological and Environmental Statistics* 3(2): 111–130. DOI: https://doi.org/10.2307/1400646.

Mugglin, A.S., B.P. Carlin, L. Zhu, and E. Colon. 1999. "Bayesian areal interpolation, estimation, and smoothing: An inferential approach for geographic information systems." *Environment and Planning* 31: 1337–1352. DOI: https://doi.org/10.1068/a311337.

OpenStreetMap Foundation. 2019. *Service Map*. Available at: https://servicemap.hel.fi/?municipality=helsinki&_rdr=Default.aspx (accessed January 2019).

Patterson, Z., M. Kryvobokov, F. Marchal, and M. Bierlaire. 2010. "Disaggregate models with aggregate Two UrbanSim applications." *Journal of Transport and Land Use* 3(2): 5–37. DOI: https://doi.org/10.5198/jtlu.v3i2.113.

Rendtel, U. and M. Ruhanen. 2018. "Die Konstruktion von Dienstleistungskarten mit Open Data am Beispiel des lokalen Bedarf an Kinderbetreuung in Berlin." *AStA Wirtschafts- und Sozialstatistisches Archiv* 12(3–4): 271–284. DOI: https://doi.org/10.1007/s11943-018-0235-y.

Schmid, T., F. Bruckschen, N. Salvati, and T. Zbiranski. 2017. "Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in Senegal." *Journal of the Royal Statistical Society: Series A* 180(4): 1163–1190. DOI: https://doi.org/10.1111/rssa.12305.

Schulz, G. 2017. "Aufwertung und Verdrängung in Berlin – Räumliche Analysen zur Messung von Gentrifizierung." *WISTA – Wirtschaft und Statistik* 4: 287–314. Available at: https://www.destatis.de/DE/Methoden/WISTA-Wirtschaft-und-Statistik/2017/04/aufwertung-verdraengung-berlin-042017.html (accessed April 2020).

Schürmann, C., R. Moeckel, and M. Wegener. 2002. "Microsimulation of urban land use." ERSA conference papers, European Regional Science Association. August 27th–31st, 2002, Dortmund, Germany.

Senate Department for Urban Development and Housing. 2017. *06.06. population density*. 2017 edition. Available at: https://www.stadtentwicklung.berlin.de/umwelt/umweltatlas/edm606_04.htm (accessed January 2019).

Senator für Stadtentwicklung und Umwelt. 2016. *Bevölkerungsprognose für Berlin und die Bezirke 2015–2030*. Available at: https://www.stadtentwicklung.berlin.de/planen/bevoelkerungsprognose/download/2015-2030/Bericht_Bevprog2015-2030.pdf (accessed February 2019).

Spiekermann, K. and M. Wegener. 2003. "Modelling urban sustainability." *International Journal of Urban Sciences* 7(1): 47–64. DOI: https://doi.org/10.1080/12265934.2003.9693522.

Statistische Ämter des Bundes und der Länder. 2015. *Zensus Atlas*. Available at: https://atlas.zensus2011.de/ (accessed January 2019).

Trevisani, M. and A. Gelfand. 2013. "Spatial misalignment models for small area estimation: a simulation study." In N. Torelli, F. Pesarin, and A. Bar-Hen, eds. *Advances in Theoretical and Applied Statistics Studies in Theoretical and Applied Statistics*: 269–279. Berlin, Heidelberg: Springer.

U.S. Census Bureau. 2017. *Small area income and poverty estimates (saipe)*. Available at: https://www.census.gov/data-tools/demo/saipe/saipe.html?s_appName=saipe&map_yearSelector=2017&map_geoSelector=aa_c (accessed January 2019).

Wand, M.P. and M.C. Jones. 1994. "Multivariate plug-in bandwidth selection." *Computational Statistics* 9(2): 97–116.

Received March 2019
Revised October 2019
Accepted December 2019

sciendo

Journal of Official Statistics, Vol. 36, No. 2, 2020, pp. 315–338, http://dx.doi.org/10.2478/JOS-2020-0017

Controlling for Selection Bias in Social Media Indicators through Official Statistics: a Proposal

Stefano M. Iacus[1], *Giuseppe Porro*[2], *Silvia Salini*[1], *and Elena Siletti*[1]

With the increase of social media usage, a huge new source of data has become available. Despite the enthusiasm linked to this revolution, one of the main outstanding criticisms in using these data is selection bias. Indeed, the reference population is unknown. Nevertheless, many studies show evidence that these data constitute a valuable source because they are more timely and possess higher space granularity. We propose to adjust statistics based on Twitter data by anchoring them to reliable official statistics through a weighted, space-time, small area estimation model. As a by-product, the proposed method also stabilizes the social media indicators, which is a welcome property required for official statistics. The method can be adapted anytime official statistics exists at the proper level of granularity and for which social media usage within the population is known. As an example, we adjust a subjective well-being indicator of "working conditions" in Italy, and combine it with relevant official statistics. The weights depend on broadband coverage and the Twitter rate at province level, while the analysis is performed at regional level. The resulting statistics are then compared with survey statistics on the "quality of job" at macro-economic regional level, showing evidence of similar paths.

Key words: Well-being; big data; sentiment analysis; small area estimation; weighting.

1. Introduction

Nowadays, researchers have potentially more data than ever before which has led to new progress in many fields of academia, government, industry, and commerce. However, although institutions and academics once had access to all the data produced because they collected or created it, have access to a smaller fraction of the data, since these data now locked up inside private companies. This information gap between the public and private sector requires further attention, but this discussion is outside of the scope of the present work. Social Networking Sites (SNS, or "social media") represent a special case for which a vast amount of data could be potentially accessible to public research.

Especially in the context of well-being measurement, the dramatic lack of timely data may be compensated by also considering alternative sources of data. SNS are a source of large and continuous flow of information, opinions, emotions, feelings and some researchers (Kwong et al. 2012; Hofacker et al. 2016) have considered them to be the largest available focus group in the world. The opinions expressed on SNS cover a large

[1] Department of Economics, Management and Quantitative Methods, University of Milan, Via Conservatorio 7 - 20122, Milan, Italy. Emails: stefano.iacus@unimi.it, silvia.salini@unimi.it, and elena.siletti@unimi.it
[2] Department of Law, Economics and Culture, Univertity of Insubria, Via Sant'Abbondio, 12 - 22100, Como, Italy. Email: giuseppe.porro@uninsubria.it.

spectrum of topics and interests, engage people from different social strata and usually do not suffer from censorship, although some exceptions have been investigated in King et al. (2013, 2014, 2017) .

Of course, these alternative sources of data are not, by design, intended to be used for the calculation of official statistics and in most cases, they are affected by different types of bias (Couper 2013). For example, in order to appear in the SNS data collections, individuals have to take some steps or satisfy some constraints like: have internet access (only 57% of the world population are in this set), open an account on the particular SNS targeted by the researchers for their analyses, and actively use it (about 45% of the world population are active users of SNS, see Table 1).

Other limitations come from the SNS themselves. No one can guarantee that these data will always exist in the future (we have seen the rise and fall of several platforms in recent years, changes of data structures and access policies). The use of public API (Application Programming Interface) or, even worse, web-scraping to obtain the data implies some lack of knowledge about the amount and quality of the data exposed by the SNS.

Despite the limitations that will be discussed in detail also in Section 3, there is a growing literature on social media as a source of data for preparing official statistics or composite indicators (see, e.g., Struijs et al. 2014; Culotta 2014; Daas et al. 2015; Alajajian et al. 2017; Tam and Clarke 2015; Kitchin 2015; Braaksma and Zeelenberg 2015; Severo et al. 2016; Van den Brakel et al. 2017) because nontraditional data are available at higher granularity, in time and space, compared to the data collected to produce official statistics.

In this article, we propose to extract emotions from social networks (Iacus et al. 2015, 2017) with the aim of building alternative subjective/perceived well-being indicators without directly surveying social network users, but only by interpreting their conversations on the internet. This approach of "listening" rather than "asking" has the potential advantage of getting rid of the nonresponse bias typical of surveys. The high-frequency rate of the data also allows taking into account that well-being is a mix of short-term, seasonal and long-term components.

Last, but not least, this article also proposes to address the selection bias problem of SNS indicators by anchoring them to official statistics and through the application of a space-time small area estimation (SAE) model (Rao 2005; Marhuenda et al. 2013) coupled with a weighting scheme.

The article is structured as follows: Section 2 introduces a multidimensional indicator of subjective well-being drawn from Twitter data: the Subjective Well-Being Index (SWBI). Section 3 discusses our proposal to control for sampling bias in Twitter-based indicators,

Table 1. Penetration data from the We Are Social and Hootsuite's report: "Digital in 2019" (Jan 2019), available at http://wearesocial.com; Annual digital growth from January 2018 to January 2019 in brackets.

Area	Internet users	Active social medial users
Global	57%(+9.1%)	45%(+9%)
European	86%(+7.6%)	55%(+3.2%)
Italian	92%(+27%)	59%(+2.9%)

combining a weighting scheme with a times-space SAE model. Section 4 restricts the focus to the component of the SWBI aimed at measuring the "quality of job/at work" and presents the results of an application of the method proposed in Section 3. Finally, Section 5 summarizes the conclusions of this work.

2. The SWBI: a Subjective Well-Being Index from Twitter Data

Since 2009, driven by the work of the Stiglitz Commission (Stiglitz et al. 2009), a large number of well-being indices have been developed – as alternatives or complements to traditional economic indicators, such as the GDP – with different structures, considering a great variety of dimensions, and for many purposes (Fleurbaey 2009). Generally, these new indicators come from survey data that, despite all efforts (Schwarz 1999; Schwarz and Strack 1999; Kahneman and Krueger 2006), still have some methodological drawbacks (Deaton 2011; Feddersen et al. 2016).

In particular, as Deaton (2011) pointed out, surveys are a potentially biased source of information; reports of well-being coming as answers to explicit questions may be influenced by contextual elements, such as the order of the questions or simply the fact that someone is asking for a personal well-being evaluation. The result is that information from surveys, as exemplified, is often subject to response error, in addition to the well-known nonresponse bias. Furthermore, surveys are costly and this makes it difficult to obtain data with a high time frequency or an adequate space granularity.

2.1. Sentiment Analysis and Twitter Data

As described in the introduction, SNS offer a large amount of data (Pentland 2014) that can be used for research purposes, enabling a new dimension of social dynamics study, as never before. Thanks to the progress of statistical methods for big data, social scientists are now able to manage and analyze data that are large in terms of dimensionality, size and time frequency (Lazer et al. 2009; King 2011). SNS like Twitter and Facebook, to mention a couple of them, have disclosed huge amounts of textual data and science shifted from traditional *text mining* to modern *sentiment/opinion analysis* with the aim of extracting semantic content from these types of data (Iacus 2014; King 2016).

The Integrated Sentiment Analysis (iSA) algorithm (Ceron et al. 2016) has been used in this work to construct a composite index of subjective well-being that attempts to capture various aspects of individual and collective life (Curini et al. 2015; Iacus et al. 2015). iSA is a human supervised machine learning method, in which a sample of texts (training set) is then first read and manually classified by human coders, and the rest of the corpus (test set) is automatically classified by the algorithm. The supervised part is essential, in that this is the step where qualitative information can be extracted from a text without relying on dictionaries or special semantic rules, but rather on cultural, psychological and emotional interpretation. Other approaches based on user-defined dictionaries exist, but mainly focus on the concept of happiness (Bollen et al. 2011; Zhao et al. 2018). The advantage of iSA over other machine learning techniques is that it is designed to directly estimate directly the aggregated distribution of the opinions (e.g., positive, negative, neutral) without passing through the individual classification of posts in the test set. This approach vastly reduces the estimation error. Moreover, as iSA is a

sequential method, in this context of highly noised data, the size of the training set needed to reach the same accuracy of other methods is usually smaller by a factor of 10 or 20 times. The reader can refer to Ceron et al. (2016) for the technical explanation of the method and its statistical properties.

It is important to note that the Twitter posts do not belong to individuals randomly chosen from a physical population (Baker et al. 2013; Murphy et al. 2014). The reference population is the population of posts of all Twitter accounts selected in the analysis. Moreover, Twitter accounts cannot be uniquely associated with individuals and some accounts are more active than others. For these reasons, the focus of our analysis is on the total volume of the posts collected (in Italy, during the reference period) through the public Twitter "search" and "streaming" API. These API are supposed to return a random sampling of the whole Twitter database, although by combining different strategies it is possible to get more. Comparing the volumes of the tweets we analyzed with the volumes obtained through a commercial provider, we could claim an almost similar coverage. However, for an institutional player, a commercial agreement should be considered as an alternative to our approach to data collection. A further restriction applies to our data set: only geo-referenced posts, about 1–5% of all tweets, have been collected. This further selection depends on individual Twitter users' privacy settings and hence may introduce additional bias. In our experience, if the analysis is based on geo-localized tweets at province level and the estimates are then aggregated at country level, the results are similar to those obtained on the whole set of tweets (with or without geo-reference information). From this personal and limited evidence, we can speculate, without any proof, that if this bias exists, it has a limited effect when data are aggregated at country level, but this is worth a further systematic investigation. To summarize, these data are clearly subject to selection bias arising in different ways: access to the internet,Twitter usage (not all people open and write on a Twitter account), Twitter platform API subsampling, and user specific privacy settings for geo-reference information. An attempt to deal with this overall bias will be presented in Section 3.

On the other hand, the advantage of using Twitter data is that the collection of data can be done in (almost) continuous time and in a wide range of sub-regional areas (in our case the Italian provinces). Finally, instead of asking something through a web form, thanks to the human supervised qualitative analysis, it is possible to capture expressions of well-being directly from the texts.

2.2. *The Construction of the SBWI Index*

The SWBI index (Iacus et al. 2015) is a multidimensional well-being indicator whose components were inspired by the dimensions adopted by the New Economic Foundation think-tank for its Happy Planet Index (New Economics Foundation 2012).

In summary, the SWBI consists of eight dimensions that concern three different well-being areas: personal well-being, social well-being, and well-being at work. More specifically,

1. *Personal well-being is defined as:*
    - **emotional well-being**: the overall balance between the frequency of experiencing positive and negative emotions, with higher scores showing that positive feelings are felt more often than negative ones (`emo`);

- **satisfying life**: having a positive assessment of one's life overall (`sat`);
- **vitality**: having energy, feeling well-rested and healthy while also being physically active (`vit`);
- **resilience and self-esteem**: a measure of individual psychological resources, of optimism and of the ability to deal with life stress (`res`); and
- **positive functioning**: feeling free to choose and having the opportunity to do it; being able to make use of personal skills while feeling absorbed and gratified in daily activities (`fun`).

2. *Social well-being is defined as:*
   - **trust and belonging**: trusting other people, feeling treated fairly and respectfully while experiencing sentiments of belonging (`tru`); and
   - **relationships**: the degree and quality of interactions in close relationships with family, friends and others who provide support (`rel`).
3. *Well-being at work is defined as:*
   - **quality of job**: feeling satished with a job, experiencing satisfaction with work-life balance, evaluating the emotional experiences of work and work conditions (`wor`).

The tweets written in the Italian language and posted from Italy constitute the SWBI's data source, and they were acquired via the public Twitter API. As mentioned, a share of the data (around 1% to 5%) includes geo-referenced information, which allows the estimation of the SWBI at a local level. As an experiment, in Iacus et al. (2019), the SWBI index has been estimated for the Italian provinces from 2012 to 2016 and compared to the "Il Sole 24 Ore" Quality of Life index (an indicator of life quality that is yearly evaluated and published by the "Il Sole 24 Ore" economic-financial newspaper in Italy).

Please note that, as SWBI does not use individual microdata, but is based on the aggregated sentiment analysis, it should be interpreted only as an aggregate measure of the level of well-being of a society.

3. A Proposal to Control for Bias in Social Media Estimates

In this section, we propose a method that makes use of official statistics to control the selection bias induced by the use of big social network data. In addition to a brief preamble to the basic SAE models, our approach, which is based on a weighted method and the SAE model, is discussed in what follows.

3.1. General SAE Models

SAE models play an important role in sampling theory and are employed when one needs to produce estimates in areas that are smaller than those for which the survey was planned. A *direct* estimator ($\hat{y}_d$), based only on the data coming from a limited-size sample from the small area, might be very unreliable; SAE *indirect* estimators are traditionally used to overcome this issue. Among indirect estimators, the model-based estimators are obtained by an explicit regression model, where a relationship between the target variable and some covariates is assumed. Model-based estimators can be classified as *unit-level* models, when

covariates are available at the unit level, and *area-level* models, when data are available only as area aggregates. In our case, as SWBI and official statistics exist only at province or regional level, the only option available is the area-level model.

The basic area-level model is the Fay-Herriot (FH) model (Fay and Herriot 1979), which is obtained as a linear mixed model in two stages consisting of a "sampling model" and a "linking model". Let $\hat{y}_d$ be a direct estimator of μ_d, a target unknown measure in area $d = 1, \ldots, D$: in the first stage, the "sampling model" (1) represents the uncertainty due to the fact that the target measure μ_d is unobservable and instead of it, only its measure on the sample $\hat{y}_d$ is available.

$$\hat{y}_d = \mu_d + e_d \tag{1}$$

$\hat{y}_d$ is unbiased, but unreliable, due to the small observed sample; and e_d are the sampling errors, which, given the characteristic of interest in d-th area, are assumed, for model convenience, to be independent and identically distributed (i.i.d.) with known variances, $N\left(0, \sigma_d^2\right)$.

In the second stage, the "linking model" (2) the area target measures μ_d are linearly related with a vector of area-level covariates $\boldsymbol{x}$.

$$\mu_d = \boldsymbol{x}_d'\boldsymbol{\beta} + u_d \tag{2}$$

where $\boldsymbol{\beta}$ is the common regression coefficients vector, and u_d are the model errors, un-observed and typically assumed i.i.d. from $N\left(0, \sigma_u^2\right)$. Combining the two model components (1) and (2), the hnal linear mixed model is dehned as follow:

$$\hat{y}_d = \boldsymbol{x}_d'\boldsymbol{\beta} + u_d + e_d \tag{3}$$

Several extensions of this basic area model have been proposed (Rao and Yu 1994; Ghosh et al. 1996; Singh et al. 2005; Marhuenda et al. 2013). Recently, these models have also been used with big data (Porter et al. 2014; Marchetti et al. 2015; Marchetti et al. 2016; Falorsi et al. 2017), which has been suggested for use as covariates when official statistics are either missing or poor. In particular, big data are used as covariates in area-level FH models, because these data are often unit level at the unit-level due to technical problems or legal restrictions. This is the case with social media search loads, remote sensing images or human mobility tracking.

Porter et al. (2014) used Google Trends searches as covariates in a spatial FH model, while in Falorsi et al. (2017), the time series query share from Google Trends was adopted as an auxiliary variable to improve the SAE model-based estimates for regional Italian youth unemployment. Marchetti et al. (2015) and Marchetti et al. (2016) have shown that big data improve the precision of small area estimates when used together with traditional covariates (i.e., official statistics or administrative data). More specifically, Marchetti et al. (2015) used big data as covariates in an FH model to estimate poverty indicators, accounting for the presence of measurement error, due to the availability of big data on mobility, using the Ybarra and Lohr (2008) approach. It is worth mentioning that Marchetti et al. (2015) suggested making use of survey data in some way to take into account the selection bias caused by the use of big data, but did not pursue this goal. This work is an attempt to implement their idea in a systematic way.

Marchetti et al. (2016) instead, used data coming from Twitter (Curini et al. 2015) as an instrumental covariate to estimate the Italian household share of food consumption expenditures at a provincial level, that is, they exploit the correlation between the official statistics indicator and social media data at regional level to reconstruct the official statistics at sub-regional level, thanks to the granularity of the Twitter data.

Conversely to the scholars cited above, in our proposal we do not use social media data (SWBI) as a covariate in a SAE model, but as a direct measure of the target unknown variable (well-being), and adopt official statistics as covariates in the area model. Following this goal, because social media data are biased, before applying the model we endorse a weighting procedure, as discussed in the next section.

3.2. *Weighting Strategy*

Usually, the methods adopted in the literature used to address the selection bias problem when using non-representative samples (e.g., the propensity score weighting (Rosembaum and Rubin 1983) or the Heckman correction (Heckman 1979)) are based on the use of unit level data (Cooper and Greenaway 2015). This also happens with social media data when individual characteristics of social media users are available. However, in light of the recently established privacy rules (GDPR) this is an increasingly remote eventuality. Note that, for Twitter data, the individual characteristics of every single account are not accurate or even unavailable and that SWBI is calculated as an aggregated estimate at province level. Unfortunately, as we will see later on, as the official statistics are available only at regional level, we adopt a hierarchical aggregation of the data at regional level, weighted by the characteristics of provincial macro-variables. As it will be explained via an application in Section 4, the macro-variables consist of the broadband coverage and the Twitter rate at provincial level. The aim is to take into account the selection bias that comes from the fact that not all people use or can use the internet and, among those who use the internet, not all of them make use of Twitter. The Twitter rate also compensates for the difference in Twitter volumes that we observe through the different geographical areas.

In particular, in Section 4, we consider $\hat{y}_{dt}^{w}$ as the regional sampling mean, where the regional units are the weighted means of province level units, in order to overcome the nonrandom sampling structure of the data:

$$\hat{y}_{dt}^{w} = \frac{1}{\sum_{i=1}^{n_{dt}} w_{idt}} \sum_{i=1}^{n_{dt}} y_{idt} w_{idt}, \tag{4}$$

where n_{dt} is the number of provinces in region d at time t, and w_{idt} are the weights. The choice of the actual weights depends on the application at hand. In Section 4, we will give a practical example. As an estimator of the variance of Equation (4), we adopt the plug-in estimator for weighted means:

$$\sigma^2_{\hat{y}^{w}_{dt}} = \frac{1}{n_{dt}} \left[\frac{1}{\sum_{i=1}^{n_{dt}} w_{idt}} \sum_{i=1}^{n_{dt}} y_{idt}^2 w_{idt} - \left(\hat{y}_{dt}^{w}\right)^2 \right]. \tag{5}$$

3.3. The Space-Time SAE Model with Weights

Since SWBI data are available for several periods of time T and domains D, we have chosen a particular SAE model, the spatio-temporal Fay-Herriot (STFH) model proposed by Marhuenda et al. (2013), to account for time and space correlations. This extension considers the spatial correlation between neighboring areas, while simultaneously including random effects for the time periods nested within areas. Thus, for domains $d = 1, 2, \ldots, D$ and time periods $t = 1, 2, \ldots, T$, let μ_{dt} be the target unknown measure (well-being) in area d at time t. The STFH model, just as any FH model, is composed of two stages. In the first stage, the "sampling model" is defined as:

$$\hat{y}^w_{dt} = \mu_{dt} + e_{dt}, \quad e_{dt} \overset{ind}{\sim} N\left(0, \sigma^2_{\hat{y}^w_{dt}}\right), \quad d = 1, 2, \ldots, D, \quad t = 1, 2, \ldots, T, \tag{6}$$

where e_{dt} are the sampling errors that are assumed to be independent and normally distributed, and $\sigma^2_{\hat{y}^w_{dt}}$ is an estimator of the variance as defined in Equation (5).

In the second stage of the STFH model, the "linking model" is as follows:

$$\mu_{dt} + \boldsymbol{x}'_{dt}\boldsymbol{\beta} + u_d + v_{dt} \quad u_d \overset{ind}{\sim} N\left(0, \sigma^2_1\right); \quad v_{dt} \overset{ind}{\sim} N\left(0, \sigma^2_2\right), \tag{7}$$

where $\boldsymbol{x}_{dt}$ is the column vector with the aggregated values of k covariates for the d-th area in t-th period of time and β is the vector of regression coefficients; u_d are the area effects that follow a first-order spatial autocorrelation process, SAR(1), with variance σ^2_1, spatial autocorrelation parameter ρ_1 and a $(d \times d)$ proximity matrix $\boldsymbol{W}$. Specifically, $\boldsymbol{W}$ is a row-standardized matrix obtained from an initial proximity matrix $\boldsymbol{W}^I$ whose diagonal elements are equal to zero and residual entries are equal to one, when the two domains are neighbours, and zero otherwise. Normality of u_d is required for the mean squared error, but not for point estimation. Furthermore, v_{dt} represents the area-time random effects that are assumed i.i.d. for each area d; these effects follow a first-order autoregressive process, AR(1), with the autocorrelation parameter ρ_2 and variance equal to σ^2_2. Accordingly, the final proposed linear mixed model is:

$$\hat{y}^w_{dt} = \boldsymbol{x}'_{dt}\boldsymbol{\beta} + u_d + v_{dt} + e_{dt}. \tag{8}$$

Therefore, $\boldsymbol{\theta} = \left(\rho_1, \sigma^2_1, \rho_2, \sigma^2_2\right)$is the vector of unknown parameters characterizing the spatio-temporal STFH model. Following Marhuenda et al. (2013), who provided $\hat{\boldsymbol{\beta}}$, the empirical best linear unbiased estimator (EBLUE) of $\boldsymbol{\beta}$, and $\hat{u}_d$ and $\hat{v}_{dt}$, the empirical best linear unbiased predictors (EBLUPs) of u_d and v_{dt}, are both obtained by replacing a consistent estimator $\hat{\boldsymbol{\theta}}$ in the respective BLUE and BLUPs introduced by Henderson (1975). The empirical estimation $\hat{\mu}_{dt}$ under the STFH model is given by:

$$\hat{\mu}_{dt} = \boldsymbol{x}'_{dt}\hat{\boldsymbol{\beta}} + \hat{u}_d + \hat{v}_{dt}. \tag{9}$$

As in Marhuenda et al. (2013), we use parametric bootstrap to estimate the mean squared error (MSE) of the EBLUPs. The MSE is calculated as follows:

$$MSE\left(\hat{\mu}_{dt}\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{u}^b_{dt} - \mu^b_{dt}\right)^2 \tag{10}$$

where, "b" remarks that these estimation is performed with the bootstrap procedure. And

$$\mu_{dt}^{b} = \boldsymbol{x}_{dt}'\hat{\boldsymbol{\beta}} + \hat{u}_{d}^{b} + \hat{v}_{dt}^{b}. \tag{11}$$

is the empirical estimation obtained in the first step of the bootstrap procedure using the bootstrap area and time effects: $\hat{u}_{d}^{b}$ and $\hat{v}_{dt}^{b}$.

In this way, the point estimate $\hat{\mu}_{dt}$ (indirect measure of well-being) of μ_{dt} (unknown well-being) can be supplemented with Equation (10) as a measure of uncertainty.

4. An Application to the Study of Well-Being at Work

The opportunity to integrate existing information on well-being with more information with a strong subjective and perceived trait, as those provided by social networks or specifically by SWBI, is a very interesting goal. In this section, with an application to Italian context we chose to use SWBI index and official statistics to guide our proposal. In particular, in Subsection 4.1, we describe the data that we use to implement the weighted procedure and the SAE model, and in Subsection 4.2 we discuss the result obtained.

4.1. Data and Variables

The SWBI index over 24 quarters from 2012 to 2017 is available at provincial and regional level. More than two hundred million tweets, in the period of the analysis were downloaded and classified, partly manually and partly through the iSA algorithm. The tweets have been classified as $+1$ (positive), 0 (neutral), or -1 (negative). The outcome variable is the estimated proportion of $+1$'s over the proportion of $+1$ and -1 and this represents the input variable y_{idt} in Equation (4).

As the variability of the number of tweets is remarkable, both along the time and the space dimension, there is the need to take into account this diversity. The range of data extends from a minimum of 1,727 tweets in 2016-Q1 for the Basilicata region to a maximum of 2,728,640 in 2017-Q2 for the Lombardia region. (Note that Valle d'Aosta has been dropped from the analysis as, considering that it consists of a single province, the proposed approach is not applicable because for example, random effects cannot be estimated.)

In order to have a more reliable view of the SWBI data at the regional level, we use the *Twitter rate* (i.e., the ratio between the number of tweets analysed and the population size in the area in the same period). The distribution of the Twitter rate over time among the Italian regions is shown in Figure 1. The average Twitter rate is around 18% ($SD = 12.29$), with a minimum regional value higher than 9% ($SD = 4.93$) in Campania, and a maximum regional value higher than 30% ($SD = 21.15$) in Molise (time averages for all the regions are blue points in the figure). The dispersion during the observational period is lower for large regions like Lazio, Puglia, Campania and Lombardia, and higher for small regions like Molise and Marche.

A better understanding of the SWBI information using the Twitter rate is made evident by examining Figure 2. The Twitter counts of 2017-Q4, shown on the left side of the figure, give the erroneous impression that most of the SWBI information comes from only a few large more populous regions (Piemonte, Lombardia, Veneto, Emilia and Campania),

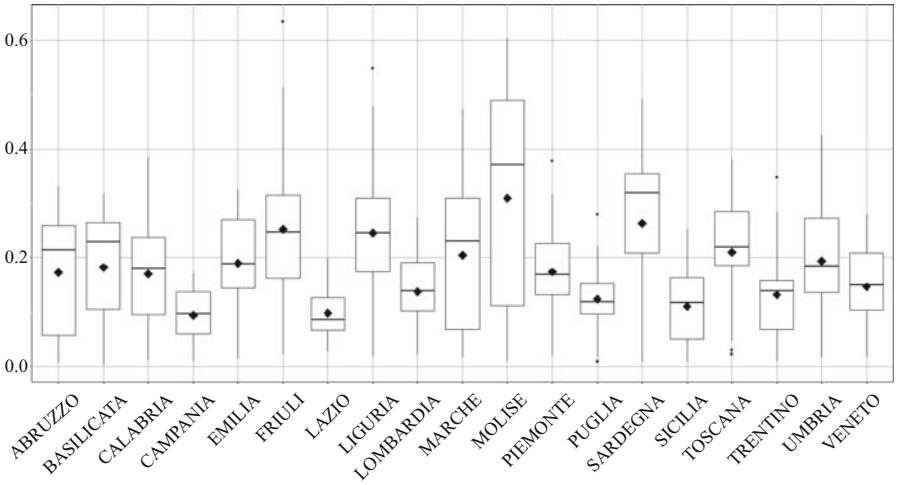


Fig. 1. Twitter rate for the considered Italian regions from the first quarter of 2012 to the last quarter of 2017.

while the Twitter rates displayed on the right side of the figure give the correct conclusion that all regions are homogeneously monitored.

4.1.1. The Construction of the Actual Weights

To implement the weighting procedure introduced in Subsection 3.2, after a selection process to define significant variables, we use the Twitter rate and the broadband coverage. The Twitter rate is closely related to mobile phone shares and broadband coverage is a measure of internet capacity. The use of these two variables is an attempt to take into account the selection bias. The Twitter rate, computed in each period and at province level,

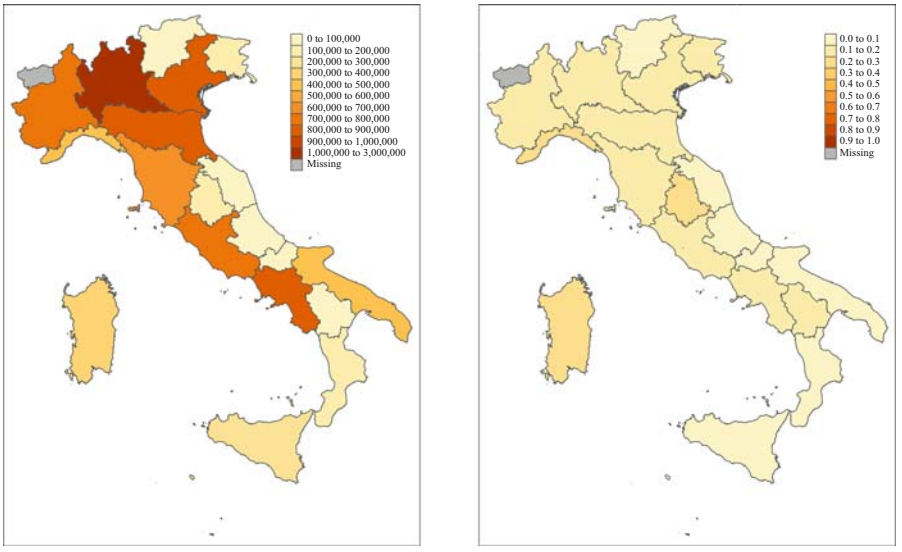


Fig. 2. Twitter counts map, on the left, and Twitter rates map, on the right, in the last quarter of 2014.

can be considered a good proxy of the use of Twitter for Italians. The broadband coverage is annual public data provided by *II Sole 24 Ore* and *Infratel Italia* for all the Italian provinces and can be considered the opportunity to access the internet in the different provinces. Coverage is quite stationary during a single year but, over time, what can happen is only an improvement of coverage in space or in signal intensity. Therefore, we replace the missing values with the data from the previous year to ensure that the coverage is not overestimated. The average broadband coverage is around 94% ($SD = 4.68$), with a minimum regional value of 72% ($SD = 4.57$) for Isernia in the Molise region. In 2012, the coverage mean was 92.15% ($SD = 3.9$) and in 2017, it was 92.65% ($SD = 5.6$). So, during the examined time period, the average broadband coverage remained the same, but the variability among regions increased, with an growth of around 42%. In detail, calling $w_{1;idt}$ the Twitter rate and $w_{2;idt}$ the broadband coverage, to apply to the weighting procedure for $\hat{y}^{w}_{dt}$ in Equation (4) and for $\sigma^{2}_{\hat{y}^{w}_{dt}}$ in Equation (5), we computed the weights as $w_{idt} = w_{1,idt} \cdot w_{2,idt}$.

4.1.2. Choosing the Covariates Among the Available Official Statistics

To apply the model proposed in Subsection 3.3, we need official statistics to use as covariates. After the Stiglitz's Commission suggestions, the Italian scenario of well-being measurement has increasingly changed. For example, the Italian National Institute of Statistics (ISTAT) set up the equitable and sustainable well-being project, where they plan a very complex system of well-being indicators, just following the same Commission suggestions. In 2013, they provided the BES ("Benessere Equo e Sostenibile", which, in English, is "Fair and Sustainable Well-being") index for the Italian regions, which analyses several dimensions of well-being.

Among these, the "work and life balance" dimension is the one that more closely relates to our research, although the construction of the composite indicator changed over time and it is not available for all quarters and provinces of Italy, making it impossible to use in our study.

ISTAT also provides other measures of well-being from the sample survey "Aspect of daily life"; however, these indicators are annual and representative for the five Italian macroeconomic areas: North-East, North-West, Center, South and Islands.

Discarding the idea to use the BES indexes and the "Aspect of daily life" survey measures, as covariates, we decided to rely on the only official statistics distributed by ISTAT that are available at least at the regional level and for the period of the analysis (although only for every quarter, the ISTAT data are available: http://dati.istat.it/ and http://demo.istat.it/). Despite the fact that the proposed model should work for each component of the SWBI at the province level, due to the limited availability of official statistics at frequencies higher than the year and at the sub-national level, we restrict our empirical analysis to the `wor` dimension of the SWBI. Even though the `wor` dimension could be monitored daily at province level, for the analysis they have been aggregated quarterly for each province ($\hat{y}_{idt}$).

The distribution of the unweighted `wor` ($\hat{y}_{dt}$) with regional aggregation over time is shown in Figure 3. The average of `wor` is 35.34% ($SD = 25.40$) with a minimum average regional value around 33% ($SD = 21.01$) in Sardegna and a maximum average regional value higher than 38% ($SD = 28.48$) in Lazio. The minimum and the maximum values of the quality of work are 9.01% for Lombardia in 2012-Q2 and 93.01% for Trentino in

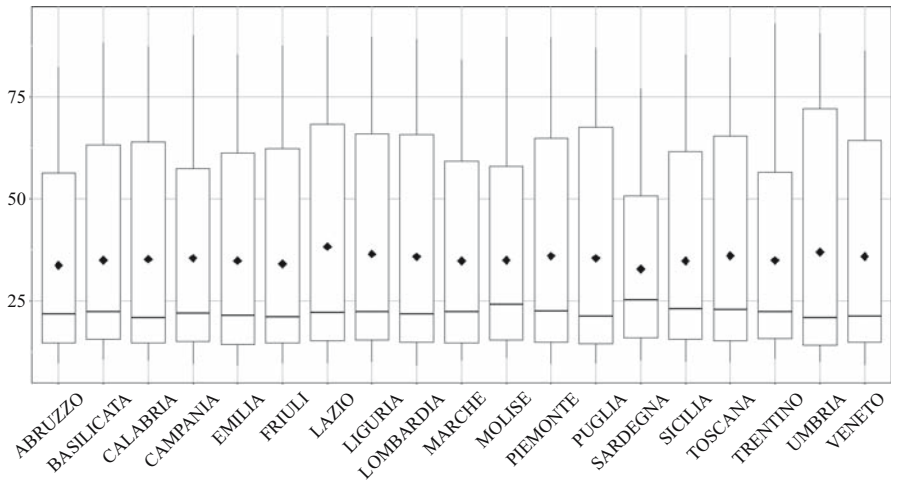


Fig. 3. The SWBI's unweighted wox *dimension ($\hat{y}_{dt}$) for the considered Italian regions from the first quarter of 2012 to the last quarter of 2017.*

2015-Q3, respectively. The similar averages are 35.79% ($SD = 26.87$) and 34.88% ($SD = 24.74$), respectively.

The considered area level auxiliary variables, before any process of selection, in the job context were as follows: the unemployment and inactivity rates, computed both in relation to the labour force (as they are traditionally calculated) and to the resident population; and the birth rate, the mortality rate and the natural rates, in the socio-demographic context. In the numerator of the natural rate there is the natural balance, which is the difference between births and deaths. After fitting the model, the selected covariates that make up the matrix $\boldsymbol{x}$ in Model (8), are the "unemployment rate" $\boldsymbol{x}_1$ and the "mortality rate" $\boldsymbol{x}_2$. The selection of these variables is the result of a standard model selection procedure after testing different variable configurations.

A large number of studies – since Clark and Oswald (1994) – provides documentary evidence of the negative relationship between unemployment and subjective well-being. It has also been argued that getting unemployed people back to work can do more for their well-being perception than subsidizing their unemployment status (see, e.g., Winkelmann 2014). In other words, non-pecuniary costs of unemployment are significant: therefore, higher unemployment rate (i.e., a higher risk of being unemployed) is here assumed to be related to the evaluation of well-being at work.

The relationship between working conditions and subjective well-being is often mediated, in the same literature, by health conditions: mortality or morbidity rates are assumed, in this respect, as proxies of health conditions.

The distribution of the unemployment rate over time among regions, as shown in Figure 4, reveals an average unemployment rate of 12.37% ($SD = 5.31$), with a minimum average regional value around 5% ($SD = 0.78$) for Trentino and a maximum average regional value higher than 22% ($SD = 2.13$) for Calabria. The same two regions also register the minimum and maximum values for the unemployment rate, 3.59% in 2017-Q3 and 25.15% in 2017-Q4, respectively.

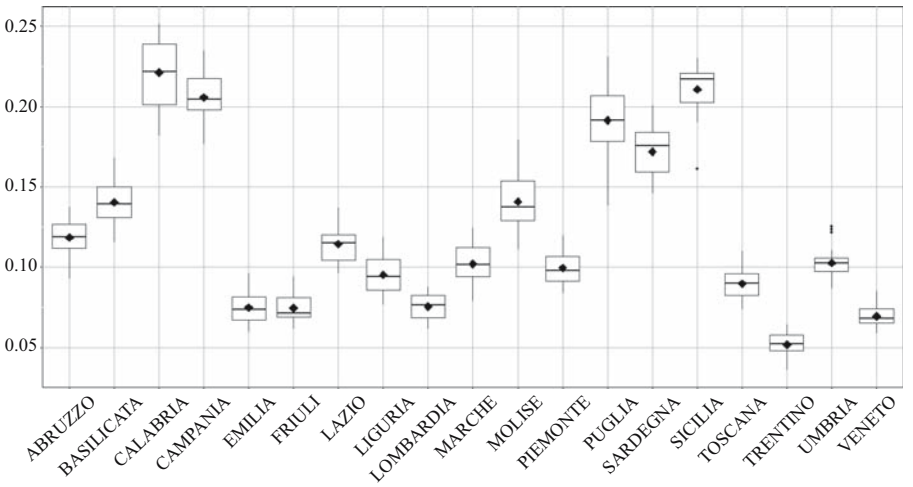


Fig. 4. Unemployment rate ($\mathbf{x}_1$) for the considered Italian regions from the first quarter of 2012 to the last quarter of 2017.

The distribution of the mortality rate over time among regions, as shown in Figure 5, illustrates an average mortality rate of 0.267% ($SD = 0.04$) with a minimum average regional value around 0.216% ($SD = 0.022$) in Trentino and a maximum average regional value higher than 0.343% ($SD = 0.032$) in Liguria. The same two regions also register the minimum and maximum values for the mortality rate, 0.19% in 2014-Q3 and 0.42% in 2017-Q1, respectively.

4.2. Results and Discussion

The weighted quality of job dimension $\hat{y}^w_{dt}$ (weighted `wor`), obtained following Equation (4), has remained stable with little variability between regions (Figure 6). The distributions

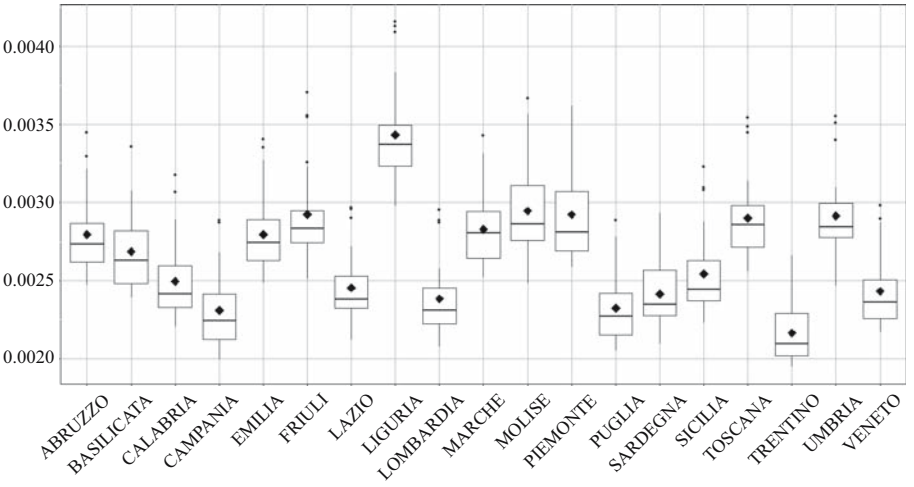


Fig. 5. Mortality rate ($\mathbf{x}_2$) for the considered Italian regions from the first quarter of 2012 to the last quarter of 2017.

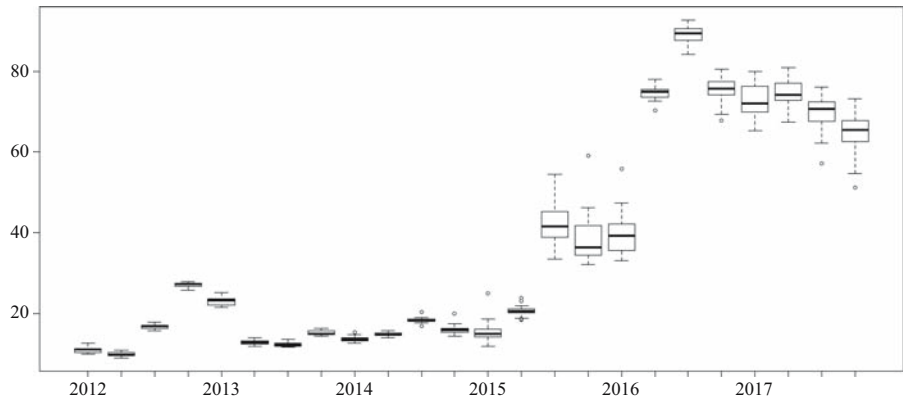


Fig. 6. The SWBI's weighted wor *dimension ($\hat{y}^w_{dt}$) during the periods from the first quarter of 2012 to the last quarter of 2017.*

were compressed until the second half of 2015, when they grew. This is especially evident from the second half of 2016, when this dimension attained values greater than 80, and even the differences between the regions were more marked, and the box-plots less crushed. Moreover, the average of $\hat{y}^w_{dt}$ is 36.17% ($SD = 26.38$) with a minimum average regional value around 34% ($SD = 22.91$) for Sardegna and a maximum average regional value higher than 39% ($SD = 29.24$) for Lazio, reflecting the earlier distributions shown in Figure 3 for $\hat{y}_{dt}$ (unweighted wor). The minimum and maximum values of the $\hat{y}^w_{dt}$ remained with Lombardia in 2012-Q2 (8.99%) and Trentino in 2015-Q3 (92.76%), respectively, and their averages were still similar (36.68% with $SD = 27.46$ for Lombardia and 37.99% with $SD = 28.66$ for Trentino).

Since comparing rankings is a valuable tool for policy makers and analysts, here we propose some discussions about them. The different rankings obtained by the two indices, both unweighted $\hat{y}_{dt}$ and weighted $\hat{y}^w_{dt}$, show no differences for around 4% of the cases ($\Delta =$ ranking differences), and only 15.6% of the cases show a Δ greater than four positions. The mean of the Δ is equal to 2.19 ($SD = 2.58$). Regions with the greatest differences were Trentino, Campania, Marche, and Sardegna, with the first two showing position improvement and the last two showing position weakening. For Trentino in particular, we remark that, after the weighting procedure, the greatest improvement took place during all four quarters of 2017.

In the applied STFH Model (8), data are available for $T = 24$ time instances, and the domains are $D = 19$, the considered Italian regions. Our data are "balanced" in that each region is measured using the same number of times and on the same occasions.

The row-standardized proximity matrix $\boldsymbol{W}_c$ of dimension 19×19 has been obtained from an initial proximity matrix, $\boldsymbol{W}^I_c$, whose diagonal elements are equal to zero and residual entries are equal to one, when the two regions had some common borders, and zero otherwise. Since in Italy, there are two regions corresponding to two islands (Sicilia and Sardegna), for these regions, we take other Italian regions with direct naval connections as neighbours.

As shown in Table 2, the coefficients for the covariates ($\hat{\beta}_1$ and $\hat{\beta}_2$) were both negative. This means that regions with larger unemployment and mortality rates had a poorer quality

Table 2. STFH model results.

(a) Estimated regression coefficients $\hat{\boldsymbol{\beta}}$ in Equation (9)

Variable	Coeff.	Std. Error	p-value
Intercept	62.72	5.49	0.000
Unemployment rate	−82.63	31.11	0.006
Mortality rate	−5649.48	1450.95	0.000

(b) Estimated vales for the vector of predictors $\hat{\boldsymbol{\theta}}$ and goodness of fit measures

Parameter	Estimate	Std. Error
$\hat{\sigma}_1^2$	0.0000	0.0000
$\hat{\rho}_1$	−0.0652	0.0000
$\hat{\sigma}_2^2$	94.72	0.0000
$\hat{\rho}_2$	0.8848	0.0000
Goodness of fit		
loglike	−1718.05	
AIC	3450.10	
BIC	3478.95	

of job dimension. The estimated spatial autocorrelation coefficient $\hat{\rho}_1$ is significant enough with a small negative value of about −0.07, (the size of the vector used is not large, $D = 19$), while the temporal autocorrelation coefficient $\hat{\rho}_2$ is still significant and has a greater positive value equal to about 0.88. The value equal to zero for $\hat{\sigma}_1^2$ is coherent with the analysis of distribution discussed above. The quality of job changes over time, but either little or not at all between regions.

4.2.1. The Weighted Measure of Well-Being at Work

In Figure 7, the scatter plots between the resulting $\hat{\mu}_{dt}$, obtained by fitting the STFH model, and the direct estimates, both unweighted $\hat{y}_{dt}$ (on the left) and weighted $\hat{y}_{dt}^w$ (on the right). In the SAE context, this graphical representation is used to test if the estimates are design unbiased: if the points lie along the diagonal, the direct estimates are approximately design unbiased, but if the points are under the line, the direct estimators are larger than the values predicted by the model, and vice versa if the points are above the line. Both the plots in the figure show points that lie along the diagonal for most of the cases. On the left side of the figure, we compare the SAE estimates $\hat{\mu}_{dt}$ with $\hat{y}_{dt}$, the unweighted estimates of `wor`, and there are more points away from the diagonal line than when the same estimated values are compared with $\hat{y}_{dt}^w$, the weighted estimates. Looking at the same plots, but for the different considered quarters, we find that the points away from the diagonal are in the periods where we have fewer analyzed tweets, and we observe an anomalous value of the variances. These two situations are caused by a lack of reliability in the information, but overall, we can conclude that the weighted estimates $\hat{y}_{dt}^w$ are approximately design unbiased.

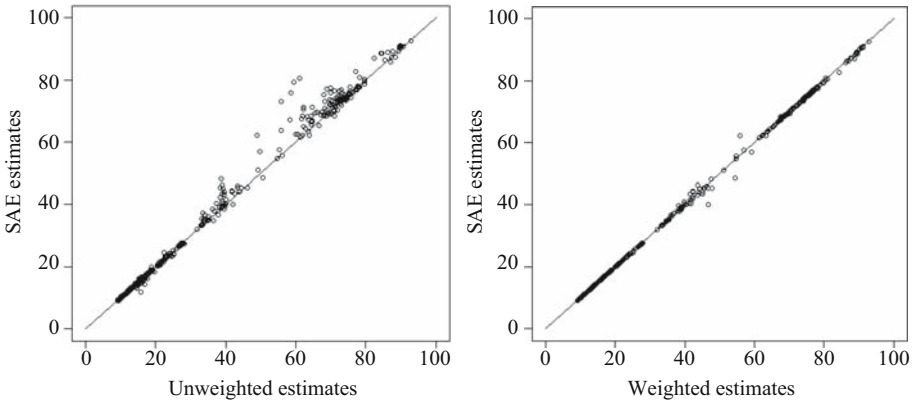


Fig. 7. Predicted values from the STFH model $\hat{\mu}_{dt}$ (SAE estimates) versus estimates of WOR, unweighted $\hat{y}_{dt}$ on the left and weighted $\hat{y}^{w}_{dt}$ on the right.

4.2.2. The Estimated Measure of Well-Being at Work from the Model

Considering the rankings, what changes if we use SAE model estimates instead of direct estimates, whether weighted or not?

Comparing the rankings obtained with the $\hat{y}_{dt}$ and those obtained with $\hat{\mu}_{dt}$,we find that in 29.2% of the cases the position is the same and in 15.8% of the cases, the Δ is greater than four. The mean of the ranking Δ is 2.16 ($SD = 2.58$). Equally, when we compared the above simple means $\hat{y}_{dt}$ with the weighted means $\hat{y}^{w}_{dt}$, regions with the greatest differences are Trentino, Campania, Marche, and Sardegna, with the first two showing position improvement and the last two showing position weakening. For Trentino, there is a great improvement during all quarters of 2017. Comparing the rankings obtained with the weighted values $\hat{y}_{dt}$ and those obtained with model estimates $\hat{\mu}_{dt}$ shows a very different situation: in 84.9% of the cases the positions are identical with less than 1% of the cases having a Δ greater than four (just one case has a great ranking difference: Marche in 2015-Q3 with a lag equal to eight positions). The average of the Δ equals 0.2 ($SD = 0.6$), which means that moving to weighted estimates $\hat{y}^{w}_{dt}$ with model predictions $\hat{\mu}_{dt}$ provides estimates that rank the same.

In SAE literature (Molina and Marhuenda 2015), coefficients of variations (CVs) are used traditionally to analyze the gain of efficiency for model estimates. While national statistical institutes are committed to publishing statistics with a high level of reliability, it is generally considered that estimates with CVs greater than 20% are not reliable. In Figures 9 and 10, the CVs of the three compared indices are shown, for the proposed final STFH model, and CVs were obtained by using the bootstrap procedure for the MSE estimates in Equation (10). As is evident, in our application, the CVs are always lower than 20%, except for fewer peaks. In particular, for most regions, the CVs are lower than 10% (Figure 10), while peak values are obtained in only a few quarters for 13 regions: Calabria, Campania, Emilia, Friuli, Lazio, Liguria, Marche, Molise, Piemonte, Lombardia, Sicilia, Toscana and Trentino. We stress that these high values of CVs are not stationary for these regions and it is clear that whenever we observe a peak of CVs, both the weighted indices and the model estimates improve reliability. Furthermore, CVs obtained for the model estimates ($\hat{\mu}_{dt}$, solid line) are always lower than the weighted estimations ($\hat{y}^{w}_{dt}$, dashed line)

Table 3. Pearson correlation coefficients r between ISTAT's WS and SAE-wor, in the five Italian geographical areas

Area	Overall	North-west	North-east	Central	South	Islands
r	0.245	0.694	0.383	0.581	0.849	0.480

and the unweighted estimates ($\hat{y}^w_{dt}$, dotted line). (For model estimates are computed as $CV = 100 \times \frac{\sqrt{MSE}}{Index}$, while for the others are $CV = 100 \times \frac{\sqrt{\text{Variance}}}{Index}$.)

Thus, values based on a STFH model look less variable in terms of the CV.

4.2.3. The Comparison Between the Estimated Measure of Well-Being at Work from the Model and an Official Index

In this section we compare our index obtained by the STFH model with an index of work satisfaction (WS) provided by ISTAT in its "Aspects of daily life" report. (All the details about the probability sample for the ISTAT survey "Aspects of daily life" can be found at www.istat.it/it/archivio/91926.)

The ISTAT's sample survey "Aspects of daily life" forms part of an integrated system of social surveys – The Multi-purpose Surveys on Household – and collects fundamental information on Italian individual and household daily life. It provides information on citizens' habits and the problems they face in everyday life. In the questionnaire, there are several thematic areas, based on different social aspects, that help describe the quality of individuals life, the degree of satisfaction of their conditions, their economic situation, the area in which they live, and the functioning of all public utility services, all topics

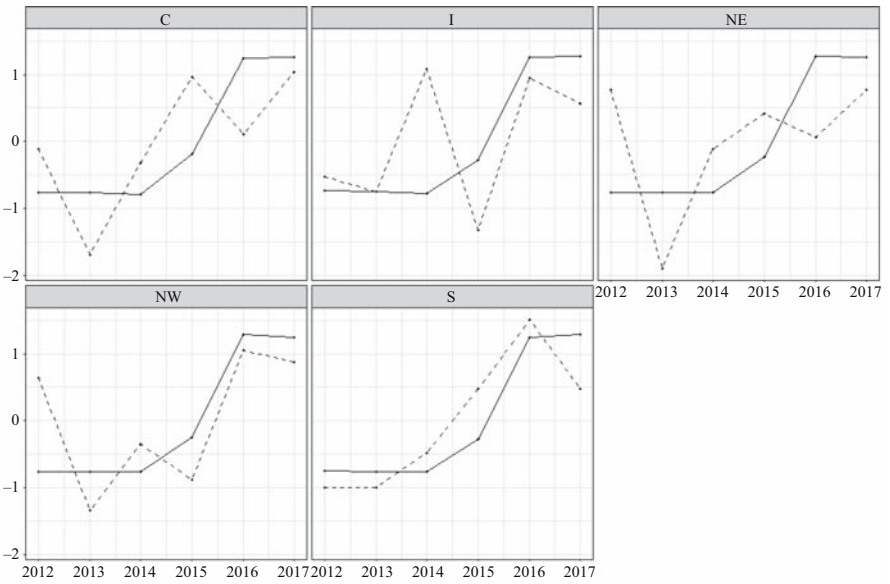


Fig. 8. Standardized time series of SAE-wor, solid line, and ISTAT's WS, dotted line, in the five Italian geographical areas (C: Central, I: Islands, NE: North-east, NW: North-west, S: South).

traditionally useful in studying the quality of life. This has been an annual survey since 2005, with data collection in February.

For our purpose we only consider WS, defined as the percentage of employed persons aged 15 years and over with a "good" level of satisfaction with their work. This index is computed as the sum of the percentages of people declaring to be "quite" and "very much" satisfied during the survey. Yearly WS data are distributed free of charge, but, as

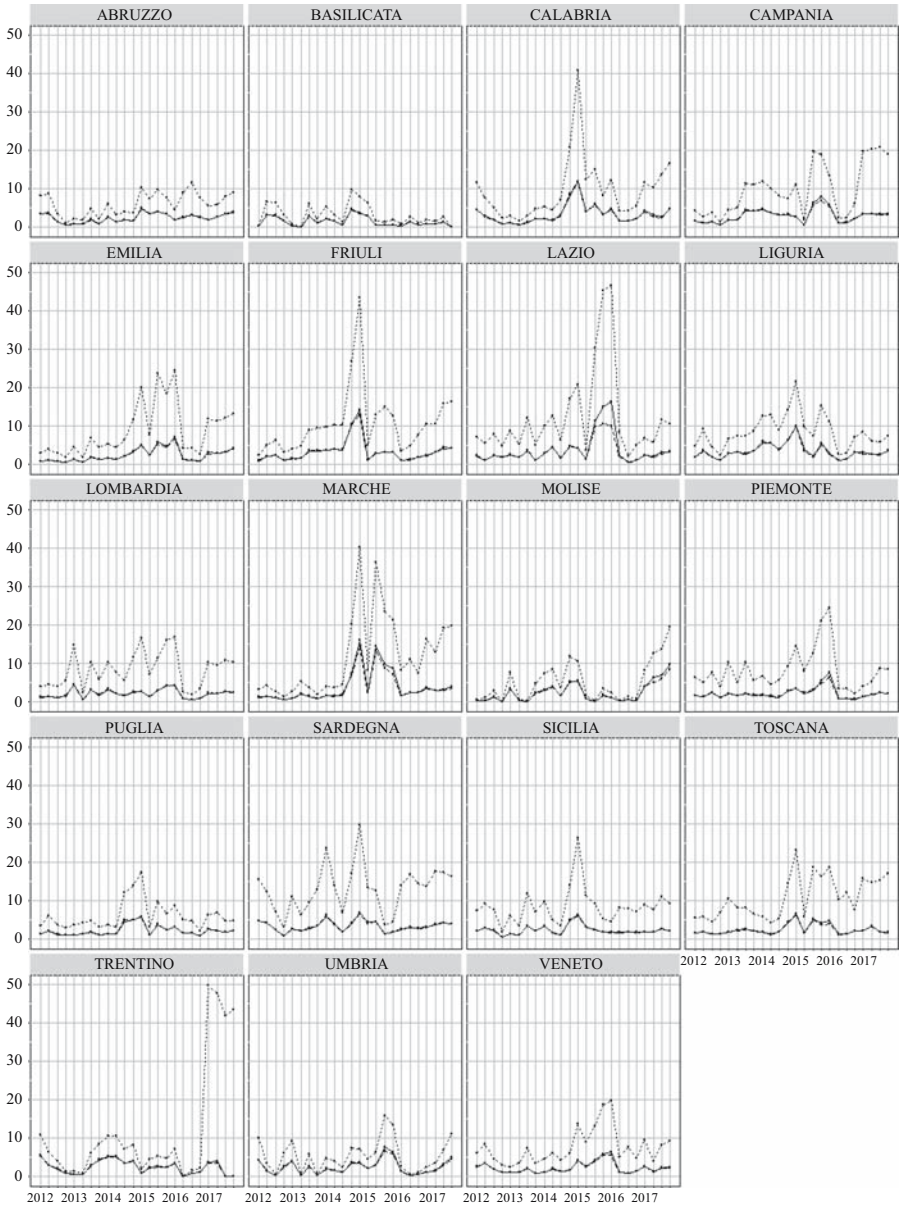


Fig. 9. Coefficient of variations for all the regions[2]; SAE estimates ($\hat{\mu}_{dt}$) with solid lines, weighted estimates ($\hat{y}^{w}_{dt}$) with dashed lines, and unweighted estimates ($\hat{y}_{dt}$) with dotted lines.

mentioned previously in the covariates section, they are representatives for the five Italian geographical areas: North-west, North-east, Central, South, and Islands.

To compare this index with our information, we aggregate the SAE estimates, $\hat{\mu}_{dt}$, obtained as discussed in the previous sections, yearly and in the same geographical areas, weighing with the corresponding resident population (SAE-`wor`).

The correlations between ISTAT index and SAE-`wor` are displayed in Table 3. If we consider all the overall data, the correlation is about 25%, while if we analyze the relationships within each area we find stronger links, with a maximum value in South Italy amounting to 85%.

Given the different scales of the ISTAT index and the proposed STFH estimator, for the purpose of visual comparison, Figure 8 represents the plot of their values, both standardized. Looking at these plots, the correlations become quite evident. We note that the correlation results are similar if we replace the STFH estimator with the raw `wor` measures (unweighted $\hat{y}^{w}_{dt}$ and weighted $\hat{y}^{w}_{dt}$).

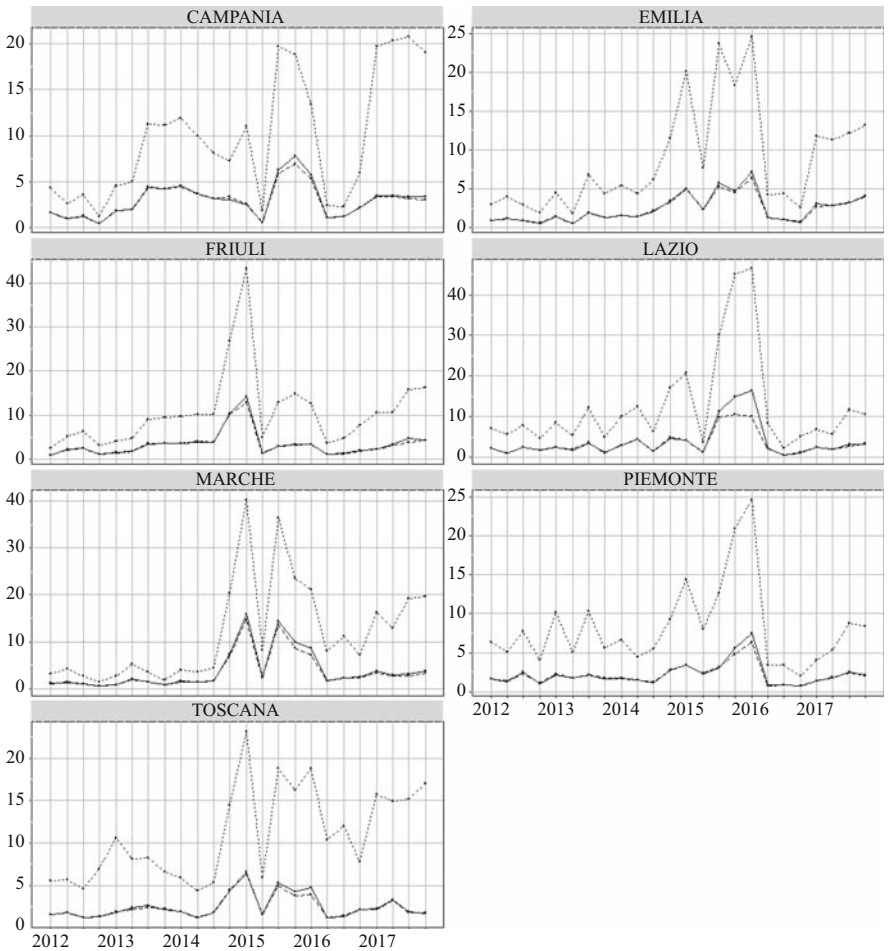


Fig. 10. Coefficient of variations for the regions with peaks greater than 20%[2]; SAE estimates ($\hat{\mu}_{dt}$) with solid lines, weighted estimates ($\hat{y}^{w}_{dt}$) with dashed lines, and unweighted estimates ($\hat{y}_{dt}$) with dotted lines.

5. Conclusion

The huge and increasing amount of data provided by social media is affected by selection bias that occurs either because not everyone has access to the internet or because not everyone who accesses the internet is interested in using social media. So far, this is a serious obstacle to using data from SNS for integrating into official statistics. To the best of our knowledge, there has been no systematic attempt to treat the bias problem, although we mentioned other important studies in which social media data have been considered along with official statistics and showed the added value in using this type of data.

In this article we have proposed to control selection bias caused by the use of aggregated data from social media by combining a weighting method and an SAE model.

Looking at the results, it seems that the selection bias inherent in social network data can be controlled using our approach. In particular, what we have shown is that – properly weighting statistics based on social media – we have approximately design unbiased statistics, that is, we have corrected the selection bias up to the only benchmark data available, which are the official statistics. We also gained additional properties through the SAE model, one of which is the stabilization of the variances of the social media statistics, which is a property required by official statistics. We have also shown that, despite using SNS data, the adjusted "wor" component of SWBI (albeit built upon different official statistics) correlates with the ISTAT statistics (available at macroeconomic level only) on the quality of work survey data.

This is clearly just the beginning of the story. Certainly, the accuracy of the proposed method could be improved using different SAE models based on dynamic systems so as to exploit fully the high resolution of the social media data, or by integrating more big data, sources at the same time, each with its own bias corrected statistics. These kinds of extensions represent interesting methodological challenges for the future.

6. References

Alajajian, S.E., J.R. Williams, A.J. Reagan, S.C. Alajajian, M.R. Frank, L. Mitchell, J. Lahne, C.M. Danforth, and P.S. Dodds. 2017. "The Lexicocalorimeter: Gauging public health through caloric input and output on social media." *PLOS ONE* 12(2)(February): 1–25. DOI: https://doi.org/10.1371/journal.pone.0168893.

Baker, R., J.M. Brick, N.A. Bates, M. Battaglia, M.P. Couper, J.A. Dever, K.J. Gile, and R. Tourangeau. 2013. "Summary Report of the AAPOR Task Force on Non-probability Sampling." *Journal of Survey Statistics and Methodology* 1(2): 90. DOI: https://doi.org/10.1093/jssam/smt008.

Bollen, J., B. Gonçalves, G. Ruan, and H. Mao. 2011. "Happiness is Assortative in Online Social Networks." *Artif. Life* (Cambridge, MA, USA) 17(3)(August): 237–251. DOI: https://doi.org/10.1162/artl_a_00034.

Braaksma, B. and K. Zeelenberg. 2015. "Re-make/Re-model: Should big data change the modelling paradigm in official statistics?" *Statistical Journal of the IAOS* 31(2): 193–202. DOI: https://doi.org/10.3233/sji-150892.

Ceron, A., L. Curini, and S.M. Iacus. 2016. "iSA: A fast, scalable and accurate algorithm for sentiment analysis of social media content." *Information Sciences* 367–368: 105–124. ISSN: 0020-0255. DOI: https://doi.org/10.1016/j.ins.2016.05.052.

Clark, A.E. and A.J. Oswald. 1994. "Unhappiness and Unemployment." *Economic Journal* 104(424): 648–659. DOI: https://doi.org/10.2307/2234639.

Cooper, D. and M. Greenaway. 2015. *Non-probability Survey Sampling in Official Statistics*. Office for National Statistics – Methodology Working Paper Series N4. Available at: https://www.k/ons/guide-method/method-quality/specific/gss-methodology-series/ons-working-paper-series/mwp3-non-probability-survey-sampling-in-official-statistics.pdf (accessed May 2020).

Couper, M.P. 2013. "Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys." *Survey Research Methods* 7(3): 145–156. ISSN: 1864-3361. DOI: https://doi.org/10.18148/srm/2013.v7i3.5751.

Culotta, A. 2014. "Estimating County Health Statistics with Twitter." In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems,* 1335–1344. CHI '14. Toronto, Ontario, Canada: ACM. ISBN: 978-1-4503-2473-1. DOI: https://doi.org/10.1145/2556288.2557139.

Curini, L., S. Iacus, and L. Canova. 2015. "Measuring Idiosyncratic Happiness Through the Analysis of Twitter: An Application to the Italian Case." *Social Indicators Research* 121(2): 525–542. ISSN: 1573-0921. DOI: https://doi.org/10.1007/s11205-014-0646-2.

Daas, P.J.H., M.J. Puts, B. Buelens, and P. A.M. van den Hurk. "Big Data as a Source for Official Statistics." *Journal of Official Statistics* 31(2): 249–262. DOI: https://doi.org/10.1515/jos-2015-0016.

Deaton, A. 2011. "The Financial Crisis and the Well-Being of America." In *Investigations in the Economics of Aging,* edited by David A. Wise, 343–368. University of Chicago Press, June.

Falorsi, S., A. Fasulo, A. Naccarato, and M. Pratesi. 2017. *Small Area model for Italian regional monthly estimates of young unemployed using Google Trends Data*. 61st World Congress of the International Statistical Institute 16–21 July 2017 – Marrakech, Marocco, October. Available at: https://www.researchgate.net/publication/320554956_Small_Area_model_for_Italian_regional_monthly_estimates_of_young_unemployed_using_Google_Trends_Data (accessed May 2020).

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(366): 269–277. ISSN: 01621459. DOI: https://doi.org/10.2307/2286322.

Feddersen, J., R. Metcalfe, and M. Wooden. 2016. "Subjective wellbeing: why weather matters." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 179(1): 203–228. ISSN: 1467-985X. DOI: https://doi.org/10.1111/rssa.12118.

Fleurbaey, M. 2009. "Beyond GDP: The Quest for a Measure of Social Welfare." *Journal of Economic Literature* 47(4): 1029–1075. DOI: https://doi.org/10.1257/jel.47.4.1029.

Ghosh, M., N. Nangia, and D.H. Kim. 1996. "Estimation of Median Income of Four-Person Families: A Bayesian Time Series Approach." *Journal of the American Statistical Association* 91(436): 1423–1431. ISSN: 01621459. DOI: https://doi.org/10.2307/2291568.

Heckman, J.J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47(1): 153–161. ISSN 00129682, 14680262. DOI: https://doi.org/10.2307/1912352.

Henderson, C.R. 1975. "Best Linear Unbiased Estimation and Prediction under a Selection Model." *Biometrics* 31(2): 423–447. ISSN 0006341X, 15410420. DOI: https://doi.org/10.2307/2529430.

Hofacker, C.F., E.C. Malthouse, and F. Sultan. 2016. "Big Data and consumer behavior: imminent opportunities." *Journal of Consumer Marketing* 33(2): 89–97. DOI: https://doi.org/10.1108/JCM-04-2015-1399.

Iacus, S.M. 2014. "Big Data or Big Fail?" The Good, the Bad and the Ugly and the missing role of Statistics. *Electronic Journal of Applied Statistical Analysis: Decision Support Systems and Services Evaluation* 5(1): 4–11. DOI: https://doi.org/10.1285/i2037-3627v5n1p4.

Iacus, S.M., G. Porro, S. Salini, and E. Siletti. 2015. "Social networks, happiness and health: from sentiment analysis to a multidimensional indicator of subjective well-being." *ArXiv e-prints* Statistics – Applications (December): 1–26. Available at: 1512.01569 [stat.AP] (accessed December 2015).

Iacus, S.M., G. Porro, S. Salini, and E. Siletti. 2017. "How to exploit big data from social networks: a subjective well-being indicator via Twitter." In *SIS 2017. Statistics and data science: new challenges, new generations. Proceedings of the Conference of the Italian Statistical Society,* edited by Alessandra Petrucci and Rosanna Verde, 537–542. 28–30 June 2017, Firenze: Firenze University Press. ISBN: 978-88-6453-521-0

Iacus, S.M., G. Porro, S. Salini, and E. Siletti. 2019. "Social Networks Data and Subjective Well-Being. An Innovative Measurement for Italian Provinces." *Scienze Regionali, Italian Journal of Regional Science* Speciale (2019): 667–678. ISSN: 1720-3929. DOI: https://doi.org/10.14650/94673.

Kahneman, D. and A.B. Krueger. 2006. "Developments in the Measurement of Subjective Well-Being." *Journal of Economic Perspectives* 20(1): 3–24. DOI: https://doi.org/10.1257/089533006776526030.

King, G. 2011. "Ensuring the Data Rich Future of the Social Sciences." *Science* 331(February): 719–721. DOI: https://doi.org/10.1126/science.1197872.

King, G. 2016. "Preface: Big Data is Not About the Data!" Chap. 1 in *Computational Social Science: Discovery and Prediction,* edited by R. Michael Alvarez, 1–10. Cambridge: Cambridge University Press.

King, G., J. Pan, and M.E. Roberts. 2013. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review* 107(2): 326–343. DOI: https://doi.org/10.1017/S0003055413000014.

King, G., J. Pan, and M.E. Roberts. 2014. "Reverse-engineering censorship in China: Randomized experimentation and participant observation." *Science* 345(6199): 891–913. ISSN: 0036-8075. DOI: https://doi.org/10.1126/science.1251722.

King, G., J. Pan, and M.E. Roberts. 2017. "How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument." *American Political Science Review* 111(3): 484–501. DOI: https://doi.org/10.1017/S0003055417000144.

Kitchin, R. 2015. "The opportunities, challenges and risks of big data for official statistics." *Statistical Journal of the IAOS* 31(3): 471–481. DOI: https://doi.org/10.3233/SJI-150906.

Kwong, B.M., S.M. McPherson, J.F.A. Shibata, and O.T. Zee. 2012. "Facebook: Data mining the world's largest focus group." *Graziadia Business Review* 15: 1–8. Available at: https://gbr.pepperdine.edu/2012/11/facebook-data-mining-the-worlds-largest-focus-group/ (accessed April 2020).

Lazer, D., A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. van Alstyne. 2009. "Computational Social Science." *Science* 323(5915): 721–723. DOI: https://doi.org/10.1126/science.1167742.

Marchetti, S., C. Giusti, and M. Pratesi. 2016. "The use of Twitter data to improve small area estimates of households' share of food consumption expenditure in Italy." *AStA Wirtschafts – und Sozialstatistisches Archiv* 10(2)(October): 79–93. ISBN 1863-8163. DOI: https://doi.org/10.1007/s11943-016-0190-4.

Marchetti, S., C. Giusti, M. Pratesi, N. Salvati, F. Giannotti, D. Pedreschi, S. Rinzivillo, L. Pappalardo, and L. Gabrielli. 2015. "Small Area Model-Based Estimators Using Big Data Sources." *Journal of Official Statistics* 31(2): 263–281. DOI: https://doi.org/10.1515/jos-2015-0017.

Marhuenda, Y., I. Molina, and D. Morales. 2013. "Small area estimation with spatio-temporal Fay-Herriot models." The Third Special Issue on Statistical Signal Extraction and Filtering, *Computational Statistics & Data Analysis* 58: 308–325. ISSN: 0167-9473. DOI: https://doi.org/10.1016/j.csda.2012.09.002.

Molina, I. and Y. Marhuenda. 2015. "sae: An R package for small area estimation." *The R Journal* 7(1): 81–98. DOI: https://doi.org/10.32614/RJ-2015-007.

Murphy, J., M.W. Link, J. Childs, C. Tesfaye, E. Dean, M. Stern, J. Pasek, J. Cohen, M. Callegaro, and P. Harwood. 2014. "Social Media in Public Opinion Research Executive summary of the AAPOR task force on Emerging Technologies in Public Opinion Research." *Public Opinion Quarterly* 78(4): 788–794. DOI: https://doi.org/10.1093/poq/nfu053.

New Economics Foundation. 2012. *The Happy Planet Index: 2012 Report. A global index of sustainable well-being*. New Economics Foundation. Available at: https://neweconomics.org/uploads/files/d8879619b64bae461f_opm6ixqee.pdf (accessed August 2015).

Pentland, A. 2014. *Social Physics: how good ideas spread – the lessons from a new science*. EBL-Schweitzer. Scribe Publications Pty Limited. ISBN: 978113143.

Porter, A.T., S.H. Holan, C.K. Wikle, and N. Cressie. 2014. "Spatial Fay-Herriot models for small area estimation with functional covariates." *Spatial Statistics* 10: 27–42. DOI: https://doi.org/10.1016/j.spasta.2014.07.001.

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(4): 511–528. ISSN: 03195724. DOI: https://doi.org/10.2307/3315407.

Rao, J.N.K. 2005. *Small Area Estimation.* Wiley Series in Survey Methodology. John Wiley & Sons, January. ISBN: 9780471431626.

Rosembaum, P.R. and D.B. Rubin. 1983. "The central role of the propensity score in observational studies for causal effects." *Biometrika* 70(1): 41–55. DOI: https://doi.org/10.2307/2335942.

Schwarz, N. 1999. "Self-reports: how the questions shape the answers." *American psychologist* 54(2): 93–105. DOI: https://doi.org/10.1037/0003-066X.54.2.93.

Schwarz, N. and F. Strack. 1999. "Reports of subjective well-being: Judgmental processes and their methodological implications." In *Well-being: The foundations of hedonic psychology,* edited by D. Kahneman, E. Diener, and N. Schwarz, 7: 61–84. New York: Russell Sage Foundation.

Severo, M., A. Feredj, and A. Romele. 2016. "Soft Data and Public Policy: Can Social Media Offer Alternatives to Official Statistics in Urban Policymaking?" *Policy & Internet* 8(3)(September): 354–372. ISSN: 1944-2866. DOI: https://doi.org/10.1002/poi3.127.

Singh, B.B., G.K. Shukla, and D. Kundu. 2005. "Spatio-temporal models in small area estimation." *Survey Methodology* 31(2): 183–195. DOI: https://doi.org/10.1.1.617.1513.

Stiglitz, J., A. Sen, and J.-P. Fitoussi. 2009. *Report by the Commission on the Measurement of Economic Performance and Social Progress*. INSEE. Available at: https://www.researchgate.net/publication/258260767_Report_of_the_Commission_on_the_Measurement_of_Economic_Performance_and_Social_Progress_CMEPSP (accessed April 2020).

Struijs, P., B. Braaksma, and P.J.H. Daas. 2014. "Official statistics and Big Data." *Big Data & Society* 1(1): 1–6. DOI: https://doi.org/10.1177/2053951714538417.

Tam, S.-M. and F. Clarke. 2015. "Big Data, Official Statistics and Some Initiatives by the Australian Bureau of Statistics." *International Statistical Review* 83(3)(December): 436–448. DOI: https://doi.org/10.1111/insr.12105.

Van den Brakel, J., J. Söhler, P.J.H. Daas, and B. Buelens. 2017. "Social media as a data source for official statistics; the Dutch Consumer Conhdence Index." *Survey Methodology* 12-001-X (43): 183–210. DOI: https://doi.org/10.13140/RG.2.2.19294.64326.

Winkelmann, R. 2014. "Unhappiness and Unemployment." *IZA World of Labor* 94. DOI: https://doi.org/10.15185/izawol.94.

Ybarra, L.M.R. and S.L. Lohr. 2008. "Small Area Estimation When Auxiliary Information Is Measured with Error." *Biometrika* 95(4): 919–931. ISSN: 00063444. DOI: https://doi.org/10.1093/biomet/asn048.

Zhao, Y., F. Yu, B. Jing, X. Hu, A. Luo, and K. Peng. 2018. "An Analysis of Well-Being Determinants at the City Level in China Using Big Data." *Social Indicators Research* (October). ISSN: 1573-0921. DOI: https://doi.org/10.1007/s11205-018-2015-z.

Received March 2019
Revised July 2019
Accepted January 2020

sciendo

Journal of Official Statistics, Vol. 36, No. 2, 2020, pp. 339–360, http://dx.doi.org/10.2478/JOS-2020-0018

Exploring Mechanisms of Recruitment and Recruitment Cooperation in Respondent Driven Sampling

Sunghee Lee[1], *Ai Rene Ong*[1], *and Michael Elliott*[1]

Respondent driven sampling (RDS) is a sampling method designed for hard-to-sample groups with strong social ties. RDS starts with a small number of arbitrarily selected participants ("*seeds*"). Seeds are issued recruitment coupons, which are used to recruit from their social networks. Waves of recruitment and data collection continue until reaching a sufficient sample size. Under the assumptions of random recruitment, with-replacement sampling, and a sufficient number of waves, the probability of selection for each participant converges to be proportional to their network size. With recruitment noncooperation, however, recruitment can end abruptly, causing operational difficulties with unstable sample sizes. Noncooperation may void the recruitment Markovian assumptions, leading to selection bias. Here, we consider two RDS studies: one targeting Korean immigrants in Los Angeles and in Michigan; and another study targeting persons who inject drugs in Southeast Michigan. We explore predictors of coupon redemption, associations between recruiter and recruits, and details within recruitment dynamics. While no consistent predictors of noncooperation were found, there was evidence that coupon redemption of targeted recruits was more common among those who shared social bonds with their recruiters, suggesting that noncooperation is more likely to be a feature of recruits not cooperating, rather than recruiters failing to distribute coupons.

Key words: Respondent driven sampling; sampling hard-to-reach population; nonresponse error.

1. Introduction

Respondent driven sampling (RDS) is a new sampling method first introduced in 1997 to address the lack of feasible approaches for capturing rare, elusive and/or hard-to-reach groups (Heckathorn 1997). RDS, as a variant of snowball sampling, is entirely different than traditional sampling. In traditional sampling, researchers control the recruitment process by selecting participants in a randomized fashion from established frames and recruiting sampled participants individually and independently. Although not universal, incentives are often provided for participation as a token of appreciation (Singer 2002). On the other hand, RDS starts with a handful of participants (*seeds*) directly recruited by researchers. After collecting data from the seeds, researchers issue recruitment coupons to

[1] Institute for Social Research, University of Michigan, 426 Thompson St., Ann Arbor, MI 48104, U.S.A. Emails: sungheel@umich.edu, aireneo@umich.edu and mrelliot@umich.edu

Acknowledgments: We thank Dr. Julie Roddy, local health departments in Southeast Michigan, Dr. Minsung Kwon, Dr. Jungran Kim and research assistants (Jae-Kyung Ahn, Karen Seo, Jenni Kim, Daayun Chung, Celina Yim, Hannah Sim and Christine Kim) for their contributions to the data collection. This research was supported by the US National Science Foundation (grant number: SES-1461470) and the US National Institutes of Health (grant number: 1-R01 AG060936-01; 1-R21 AG062844-01).

them, who then distribute the coupons to and recruit their peers from their own social networks. The peers participate in the study by redeeming the coupons and, just like the seeds, are issued coupons for recruiting their own peers. This recruitment process continues in waves. Although there is no standard, RDS coupons include a serial number that links a recruit to his/her recruiter and are used as a means to incentivize recruitment (i.e., in addition to study participation incentives, participants are given recruitment incentives based on the number of redeemed coupons). The recruitment process in RDS is essentially controlled by participants, is incentivized and is based on chain-referrals, where each seed forms his/her own chain. Naturally, word of mouth (WOM) comes into play in RDS (Hathaway et al. 2010).

RDS is neither the only option nor a perfect method for reaching rare, elusive and/or hard-to-reach groups, as illustrated in Lee et al. (2014) and Wagner and Lee (2014). It depends on the study goals, as well as the characteristics of the target groups. Obviously, one may consider screening probability samples potentially stratified by characteristics related to the target rare group (Kalsbeek 2003; Kalton and Anderson 1986). However, this is resource-intensive and becomes more so when the target groups are geographically dispersed. For rare groups with stigmatized characteristics (e.g., illicit substance users), this screening method may be ineffective. If there are certain access points in the geography or cyberspace (e.g., gay dating apps, ethnic groceries) frequented by a large share of target rare groups, those points may be leveraged for sampling (e.g., time and location sampling). When the target groups are difficult to reach through conventional methods or existing venues but are connected with each other in some fashion, RDS may become effective. With a social network being the basic premise of RDS, if members of the target group are not networked, RDS is inapplicable.

2. Recruitment Cooperation in Respondent Driven Sampling

Success in implementing RDS depends on participants' cooperation with recruitment requests. Noncooperation directly leads to recruitment chains dying out and causes samples to stop growing in size. This forces researchers to make design changes to meet the target sample size. For example, an RDS study, the Chicago Health and Life Experiences of Women, reports slow data collection, forcing them to improvise design features (e.g., adding more seeds) (Bostwick et al. 2015; Martin et al. 2015). Noncooperation with RDS recruitment is compound in nature and can be attributed to four sources: 1) participants' network sizes (degree); 2) participants' willingness to recruit their peers which can be manifested through their acceptance of coupons from the researchers and/or their act of giving out coupons; 3) their peers' acceptance of coupons; and 4) coupon recipients' willingness to participate (Lee et al. 2012; Gile et al. 2015; Lee et al. 2017). Naturally, participants with small networks are likely to have fewer eligible peers than those with larger networks. This can be ascertained to some extent by directly asking the network size. The latter three sources are typically unknowable until the data collection ends. This turns RDS sample sizes into a random variable for which researchers have little information, making them neither predictable nor controllable (World Health Organization and UNAIDS 2013; Centers for Disease Control and Prevention, CDC 2015; Lee et al. 2018). Examples of such difficulties are plentiful in the field (e.g., LGBTQ in

Washington D.C. in Tucker et al. 2015; Polish migrants in Great Britain in Luthra 2011); however, they remain as anecdotes reported mostly at professional meetings and are rarely found in the peer-reviewed literature. Exacerbated by the lack of transparency and inadequate reporting of RDS studies (Hafeez 2012; White et al. 2015), when facing unforeseen challenges in RDS operations, researchers will be left on their own to make design changes on the spur of the moment in hopes of making RDS "work" (Martin et al. 2015). This approach is neither replicable nor informative for effective RDS designs.

Recruitment cooperation has profound implications for inferences. However, it is largely overlooked in the extant RDS literature which rests on strong assumptions (Lee 2009; Gile and Handcock 2010; White et al. 2015; Heckathorn and Cameron 2017; Lee et al. 2017). Arguably, the most important assumption is that RDS recruitment chains follow a Markov process, where the future and past states are independent given the present. Figures 1 (A, B) illustrates two different RDS scenarios of a sample size 20 coming from two chains. In Figure 1A, chains start with seeds' characteristics (black or white color), which are lost in the process of recruiting through ten waves, and the characteristics of the overall sample becomes independent of the seeds. On the other hand, the two chains in Figure 1B are different in their sizes and much shorter at four and two waves, with chains "remembering" the characteristics of their own seeds. Chains that are long and similar in their length and size (e.g., Figure 1A) are likely to follow a Markov process. Noncooperation, particularly differential noncooperation, produces chains like Figure 1B, violating this assumption (Strömdahl et al. 2015). This dependence needs to be accounted for in computing sampling variance. Although noted as a critical gap in RDS inferences (Heckathorn and Cameron 2017), variance estimation becomes less of a concern if the design facilitates the chains to grow like Figure 1A.

While noncooperation and sampling are conceptualized separately in traditional sampling (Groves 1989), noncooperation in RDS directly influences the sampling mechanism (Lee 2009). It is imperative for inferences to reflect design and operational glitches (Shadish and Cook 1999). For instance, in probability sampling, estimators incorporate design fully (e.g., selection probability, stratification) and attempt to correct for nonresponse based on the extensive literature on its mechanism (e.g., Groves and Couper 1998; Kalton and Flores-Cervantes 2003). While the same should hold for RDS, existing RDS inference approaches rely mostly on the network sizes and structures (McCreesh et al. 2013; Tomas and Gile 2011; Gile and Handcock 2015; Li et al. 2017), and are rather blind to the realities of recruitment noncooperation (Gile et al. 2015) despite

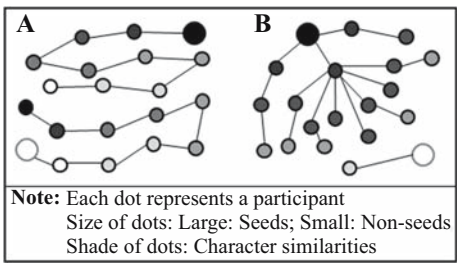


Fig. 1 (A, B). Two types of recruitment chains with two seeds.

its nonrandomness (Abramovitz et al. 2009; Lee et al. 2017) and undesirable effects (Stein et al. 2018). Hence, poor inference properties of existing RDS estimators for real-world data (Lu et al. 2012; Verdery et al. 2015; Selvaraj et al. 2016) are not surprising.

This study uses data from two RDS studies that targeted distinctive rare groups using different administration modes. These studies embedded similar features to allow us to examine the dynamics of RDS recruitment, in particular the mechanisms of recruitment noncooperation, which have not been carefully examined in the extant literature.

3. Data and Methods

3.1. Data

3.1.1. Study 1: Positive Attitudes Towards Health

We conducted the Positive Attitudes Towards Health (PATH), an in-person RDS study targeting persons who inject drugs (PWID) in Southeast Michigan in the United States, from May to November 2017.

The overall study protocol closely resembled the PWID component of the National HIV Behavioral Surveillance by the Centers for Disease Control and Prevention (CDC 2015). Seeds were recruited through recruitment flyers and cards distributed through PWID and at-risk population service agencies around Wayne, Macomb and St. Clair counties Interested individuals were instructed to call the research team, who then conducted a short screening interview and scheduled an in-person visit with eligible persons. Specifically, PWID who were at least 18 years old, who resided in the three counties noted above, and who injected within the six months were eligible. The data collection sites were Detroit, Warren, Roseville and St. Clair. Upon each visit, interviewers conducted another screening interview that included checking for physical injection marks and administered the main survey as an audio computer-assisted-self-interview (A-CASI). After the main survey, recruitment coupons were issued to participants who were also provided with instructions regarding peer recruitment for the PATH and a recruitment instruction card. Three coupons were issued per participant unless they indicated knowing fewer than three PWID in the area, in which case they were given a number of coupons up to the number of eligible people they know (i.e., one coupon if they knew one other person, two if they knew two). Reminder calls regarding recruitment were given to the participants 7, 10 and 12 days after the main survey. Roughly two weeks after the main interview, all participants to whom at least one coupon was issued were invited back to the site for a follow-up survey also done in A-CASI and for a recruitment incentive payment. Participants were offered USD 30 for completing the main survey, USD 5 for the follow-up survey and USD 10 for each successful recruit. Overall, 410 PWIDs participated in PATH, and 172 participated in the follow-up. Socio-demographic characteristics of the PATH participants are provided in Table A1 (Appendix, Section 11).

3.1.2. Study 2: Health and Life Study of Koreans

A web-based RDS was implemented through the Health and Life Study of Koreans (HLSK) which targeted Korean immigrants (i.e., Koreans born outside of the United

States) living in Los Angeles County (LA) and the State of Michigan (MI). The recruitment was done through RDS chain referrals. Unlike the PATH, the majority of the operation was done over the web. Most seeds were recruited via online ads and some referrals through various Korean and Korean American organizations. Eligible seeds were invited to the main survey by the study team and given a unique number required to access the questionnaire. Toward the end of the main survey, participants were notified about the peer recruitment. Shortly after the main survey, two coupons were issued, unless participants reported knowing fewer than two foreign-born Koreans in the target area. Coupons included unique numbers for the recruits to use for accessing questionnaires. When coupons were not redeemed, we sent reminders to the participants 7, 10 and 12 days after the main survey to encourage them to distribute coupons. Two weeks after the main survey, participants were invited to a follow-up survey. A total of 639 Koreans participated in the main survey and 266 in its follow-up survey. Note that among 639 main completes, eight were Korean immigrants whose postal addresses are outside of Louisiana and Michigan. Additionally, there were two participants whose survey responses were not properly stored in the software and, hence, were excluded from the analysis of survey responses. Table A1 (Appendix) provides socio-demographic characteristics of the HLSK participants.

3.1.3. Features Specific to Recruitment Process in Study 1 and Study 2

In order to examine dynamics of peer-referral recruitment processes in both PATH and HLSK, we implemented special features as described in Figure 2. In the main survey, all participants were asked how many study-eligible persons they knew ("social network size") and about the age, gender and relationship of the persons to whom they intended to distribute coupons (i.e., intended recruits). Specifically for non-seeds (i.e., participants recruited by someone else), we asked about the characteristics of their recruiters, including age, gender, relationship type, closeness and contact patterns. It should be noted that the social network size was examined in various ways in each survey. In addition to the standard measures that simply asked how many target group members a participant knew,

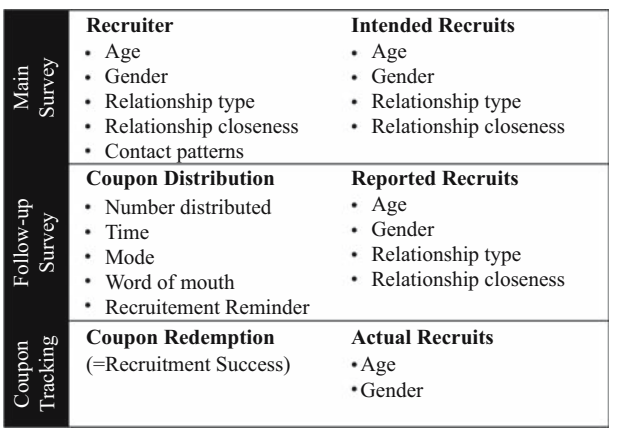


Fig. 2. Features for recruitment noncooperation examination in PATH and HLSK.

both surveys asked how many of them were participants' family members, how many of them participants interacted with more than once a week, and how many of them participants felt close to.

The follow-up survey included questions detailing the actual coupon distribution process, rather than the intended distribution. Specifically, we asked for the number of coupons each participant actually distributed; the characteristics (age, gender and relationship) of the persons to whom coupons were distributed (i.e., reported recruits); the timing and mode of coupon distribution; the type of message used when distributing coupons ("What did you tell (RECRUIT) about the study?"); whether they knew the participation status of the reported recruits; and for HLSK, utility of the recruitment reminder.

As part of the sample management, we tracked each of the issued coupons and the redemption status through which participants' recruitment success can be ascertained. From this, the links between participants and their recruiters and between participants and their actual recruits can be established. Because characteristics of actual recruits can be accessed from the main survey, they can be added to our analysis. Within each study, when the main, follow-up and the tracking data are combined, data about the characteristics of intended, attempted and actual recruits become available.

3.2. Analysis Procedures

The analysis will focus on five aspects of peer recruitment in RDS. Because similar features were implemented in PATH and HLSK, all analysis will be done using both surveys to broaden the applicability.

First, we will examine the progress of the main survey data collection by plotting the number of recruited seeds, the number of issued coupons and the number of completed interviews (i.e., participants) separately for each data collection site over time as well as visualizing the overall recruitment chain structures that denote each participant as a node and each recruiter-recruit relationship as an edge. In particular, the recruitment network diagrams will be compared to Figures 1 (A, B). Here, visualized recruitment networks will allow us to assess the plausibility of the Markov process assumption in Studies 1 and 2.

The second analysis will focus on the coupon use, including the distribution and the redemption. From coupon tracking data, we will examine the number of issued and successfully redeemed coupons. Further, the follow-up survey included questions about participants' coupon distribution and their knowledge about distributed coupons' redemption status. This allows us to examine, for the follow-up participants, whether the number of distributed coupons reported in the follow-up survey is greater than, smaller than or equal to the number of redeemed coupons. By linking knowledge about the coupon redemption status from the follow-up survey and the true redemption status ascertained from coupon tracking, we will examine how participants' knowledge compares to the actual success.

Recruitment success among all participants who were issued coupons will be examined in a multivariate model as a function of participant characteristics, including socio-demographics in Table A1 (Appendix), social network size and drug use for the PATH or ethnic identity for the HLSK. In particular, we will use the ratio of the number of redeemed

coupons over issued coupons as a dependent variable in a quasibinomial regression. This is to avoid any confounding imposed by the number of issued coupons on recruitment success and to account for overdispersion. In these models, we tested all versions of social networks described in Subsection 3.1.3 as an independent variable. The best model fit was observed through likelihood ratio tests for the following versions of social networks: for the PATH, the number of PWID that participants interacted more than once a week; and for the HLSK, the number of foreign-born Koreans to whom participants felt close. For the PATH, we also tested models with participants' drug use. This did not improve the ability to explain recruitment success. We present results from the best fitting models.

The third analysis will involve triangulating data from the main survey, the follow-up survey and coupon tracking, and examining profiles of intended (from the main survey), reported (from the main survey) and actual recruits (from coupon tracking and the main survey). We will compare profiles of all intended recruits reported by all participants and actual recruits ascertained from coupon linkage, focusing on age and sex. Intended recruits will be further compared by participants' recruitment success in order to examine whether successful recruiters target different types of peers than unsuccessful recruiters through χ^2 independence tests. For the follow-up participants, the profiles will be compared between their intended and reported recruits. Further, for the follow-up participants who also were successful at recruitment, age and sex of their intended, reported and actual recruits will be compared. There is no way to link intended, reported and actual recruits individually as we will have no data on all intended or reported recruits who did not participate. Instead, we will focus on the profile of the recruits. Note that the data on intended, reported and actual recruits will each be stacked at the recruit level in the analysis of recruit profiles.

The social relationship between recruiters and recruits will be our fourth analysis. For all participants, the main survey asked the type of relationship and the closeness that participants have with their intended recruits; and for any non-seed participants (i.e., actual recruits), the main survey also asked the relationship type and closeness with their recruiter. These allow us to compare the relationship between recruiters and their intended recruits and between actual recruits and their recruiters. As done with recruit profiles, intended recruits will be further compared by participants' recruitment success in order to examine whether social relationship matters for successful recruitment through χ^2 independence tests. For the follow-up participants, the relationship type and closeness will be compared between their intended and reported recruits.

The last analysis step will examine the dynamics within the coupon distribution process ascertained through the follow-up survey: in particular, the coupon distribution timing since issuance, and its mode and WOM participants used during coupon distribution. For the HLSK, we will also examine whether participants received recruitment reminders and whether they reminded the reported recruits about participating in the survey.

To date, there is no comprehensive analysis on RDS recruitment cooperation. Therefore, our analysis will take a descriptive approach to provide detailed information in the recruitment process that can be ascertained from our triangulated data. The next four sections will include results from each analysis step along with implications. These implications will be summarized in the last section along with comparisons of two RDS studies in our analysis.

4. Data Collection Progress

Figures 3 (A, B) include the number of seeds, issued coupons and participants (i.e., seeds and redeemed coupons) over the data collection periods of the PATH and the HLSK. The PATH started its data collection in the first week of May and continued until the first week of November 2017. A total of 410 participants (286 Detroit, 104 St. Clair, 20 Macomb) stemming from 46 seeds (22 Detroit, 14 St. Clair, 10 Macomb) were interviewed, meaning 364 were recruited via coupons. Most seeds were recruited at the beginning of the data collection, and very few were added in the second half. It is notable that, even though the number of seeds did not differ greatly across sites, the number of non-seeds did. Especially in Macomb, after two months into data collection, only ten recruits were generated from

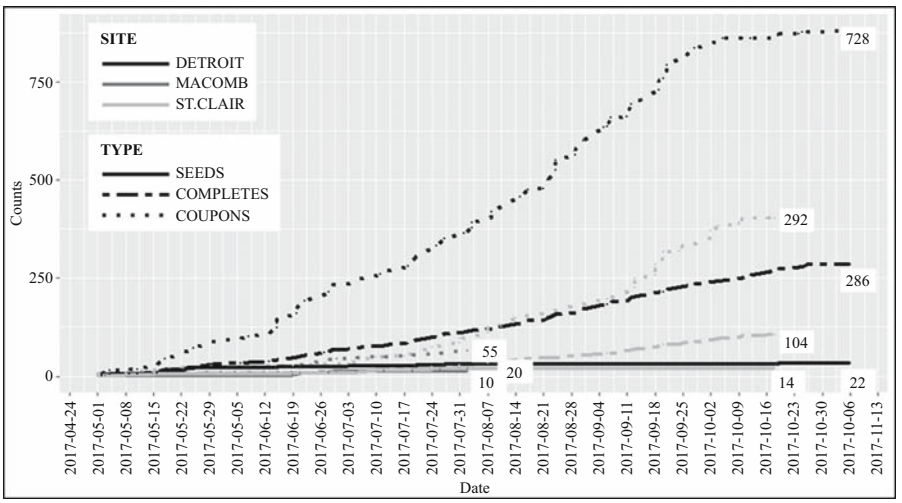


Fig. 3A. Data collection progress, positive attitudes towards health.

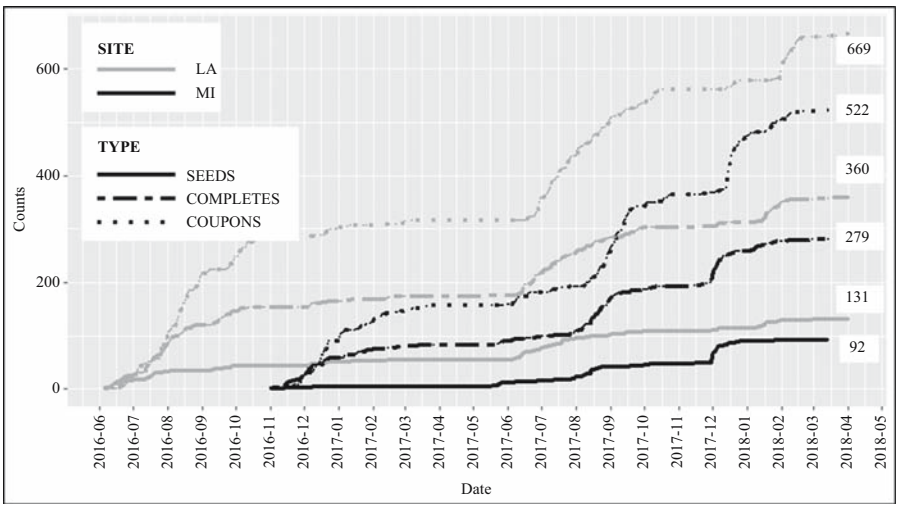


Fig. 3B. Data collection progress, health and life study of Koreans.

ten seeds with no in-person visits scheduled. Due to low productivity, data collection was suspended in Macomb as of August 2017.

The HLSK started in June 2016 with a small number of seeds in Louisiana. Michigan was added in November 2016. The data collection continued until late March 2018. From a total of 223 seeds (131 Louisiana, 92 Michigan), 411 recruits (229 Louisiana, 187 Michigan) were generated through coupons, resulting in a total sample size of 639 (360 Louisiana, 279 Michigan). Unlike the PATH, the number of completes did not grow gradually. Rather, the sample size growth plateaued at points and chains died out. To increase the number of completes, we were required to add seeds. We added seeds in batches when the data collection progress was lagging behind. This can be seen from the correspondence between counts of seeds and increases in completes in Figure 3B.

Figure 4A provides structures of all 46 chains in the PATH. There was a large variation in chain lengths and sizes. Among 46 chains, about half (24) did not generate any additional participant after seeds. The longest chain lasted for 14 waves after the seeds. The sample size per chain was distributed positively skewed with a mean of 8.9, a median of 1 and a maximum of 78. This skewed distribution of chain length and chain sample size held true across sites. Out of 223 chains in the HLSK shown in Figure 4B, nearly half (112 chains) died at the seed. The longest chain in Louisiana recruited through nine waves after seeds and in Michigan through 12 waves. More than three quarters of chains recruited three or fewer participants, but 17 Koreans were recruited from the longest chains in Louisiana and 48 in Michigan. Clearly, both recruitment chains resemble Figure 1B rather than 1A. Not surprisingly, the correlation coefficient between chain lengths and sizes in the PATH and HLSK were estimated around 0.90.

Overall, our data collection progress shows that chains grow in a way that far from meets the Markov process assumption. The sample size growth is also shown to be unpredictable. Within the same RDS survey of PWID that used the same data collection protocols, some sites were less productive than others. This resulted in a closing of the least productive site.

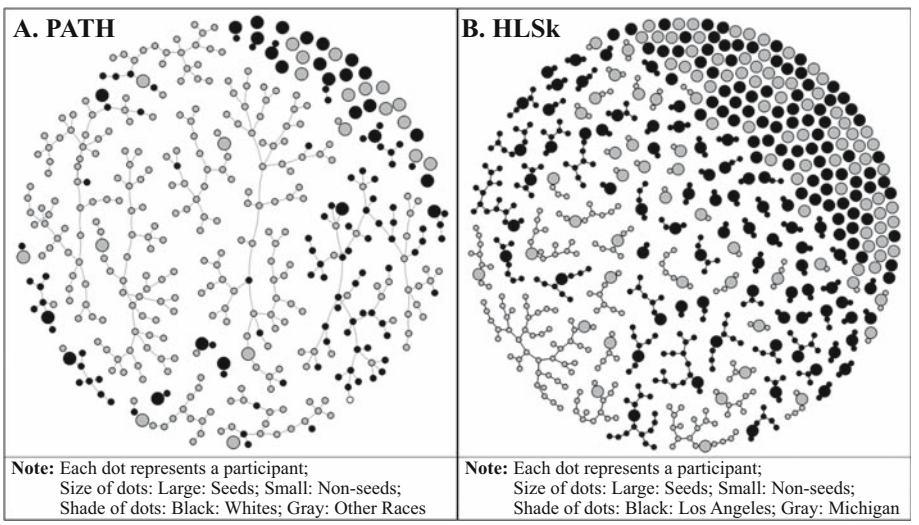


Fig. 4 (A, B). Recruitment chain graphs.

In a Web RDS study, it was necessary to add more seeds in the middle of the data collection to continue the data collection and meet the target sample size.

5. Coupon Distribution and Redemption

A total of 1,075 coupons were issued to 367 PATH participants. Among them, 364 coupons were redeemed (i.e., successful recruitment), producing a 33.9% coupon redemption rate. In the HLSK, among the 1,191 coupons issued to 607 participants, 416 were redeemed (a 34.9% redemption rate). At the respondent level, 56.9% of the PATH participants and 46.8% of the HLSK participants successfully recruited one or more peers.

In Table 1, the number of redeemed coupons at the participant level is shown in relation to the number of issued coupons. In both PATH and HLSK, the majority of participants (349 out of 410 in PATH; 584 out of 639 in HLSK) received the maximum number of coupons. However, the rate of all issued coupons being redeemed was low. In the PATH, 5.0% of those issued one coupon, 10.0% of those issued two coupons and 12.9% of those issued three coupons had all coupons successfully redeemed. In the HLSK, 43.5% of those issued one coupon and 22.6% of those two coupons had all coupons successfully redeemed. For the remaining participants, redeemed coupons were fewer than the issued.

Information about how many coupons were distributed can be ascertained for the follow-up participants. It should be noted that follow-up participation was significantly related to recruitment success as 89.5% of the PATH follow-up participants successfully recruited one or more peers, whereas 28.2% of non-participants did so ($t = 15.4$, $p < 0.001$). The corresponding rates for the HLSK were 63.2% for follow-up participants and 32.8% for nonparticipants ($t = 7.7$, $p < 0.001$). In the follow-up survey, 91.1% and 74.0% of the PATH and the HLSK participants, respectively, reported distributing the same number of coupons as issued. Table 2 compares the number of redeemed coupons ascertained from coupon tracking against the number of distributed coupons reported in the follow-up survey, as well as against the number of distributed coupons that participants knew to have been redeemed by the coupon recipient peer(s). As one would expect from Figure 3, as well as based on the principles of nonresponse, more coupons were distributed than redeemed for 65.6% of the PATH participants and for 44.5% of the HLSK participants. When using participants' knowledge about the redemption status of the coupons they distributed, only a small proportion of the participants reported not knowing the status (2.9% for the PATH; 6.2% for the HLSK). In fact, the majority of the follow-up participants' knowledge matched with the actual redemption status (64.0% for the PATH; 66.9% for the HLSK). PATH participants were more likely to underreport than overreport the redemption status (25.6% underreport versus 7.6% overreport), whereas HLSK participants were more likely to overreport than underreport (16.7% overreport versus 10.2% underreport). Table 3 includes the results of quasibinomial regression models that predicted the probability of an issued coupon being redeemed. In the PATH, age was the only predictor with marginal significance ($p = 0.070$): coupons distributed by participants in the oldest category (61 years old or older) were more likely to be redeemed than those distributed by participants aged 18–40 years old. In the HLSK, successful coupon redemption was associated with marital status, interview language and network size measured by the number of peers to

Table 1. Number of issued coupons and redeemed coupons per respondent.

	PATH					HLSK		
	No. redeemed coupons %					No. redeemed coupons %		
No. issued Coupons	0 (n = 158)	1 (n = 99)	2 (n = 65)	3 (n = 45)	No. issued coupons	0 (n = 355)	1 (n = 152)	2 (n = 132)
0 (n = 43)	-	-	-	-	0 (n = 32)	-	-	-
1 (n = 8)	50.0	50.0	-	-	1 (n = 23)	56.5	43.5	-
2 (n = 10)	70.0	20.0	10.0	-	2 (n = 584)	53.1	24.3	22.6
3 (n = 349)	42.1	26.7	18.3	12.9				

Table 2. Comparison of number of distributed coupons reported in the follow-up survey versus redeemed coupons from coupon tracking and number of distributed coupons reported as redeemed by peers in the follow-up survey versus redeemed coupons from coupon tracking.

	PATH %	HLSK %
No. coupons reported as distributed is	(n = 157)	(n = 265)
Greater than No. redeemed coupons	65.6	44.5
Equal to No. redeemed coupons	34.4	53.6
Smaller than No. redeemed coupons	-	1.9
No. coupons reported as used by the peer(s) is	(n = 172)	(n = 275)
Greater than No. redeemed coupons	7.6	16.7
Equal to No. redeemed coupons	64.0	66.9
Smaller than No. redeemed coupons	25.6	10.2
Don't know	2.9	6.2

whom participants felt close significantly at $p < 0.050$, and with age, sex, and employment status marginally significantly at $p < 0.100$. In particular, the probability of a coupon being redeemed was higher for coupons distributed by married participants than those not married; by participants who took the survey in Korean rather than English; and by participants with larger network sizes. Marginally lower coupon redemption probabilities were observed from participants who were aged 50–59 years old (ref: 18–29 years old), male or employed, compared to their counterparts.

Between the PATH and the HLSK, the redemption rate of a given coupon was around 30%. There was a systematic pattern in recruitment success as participants' certain socio-demographics as well as social network sizes were significant in predicting the probability of coupons issued to them being redeemed by their peers. Network sizes as currently measured in RDS were not effective in predicting recruitment success. Rather, network sizes that were restricted by certain social relationship (e.g., closeness) were effective for such predictions. Follow-up surveys indicated that, even though most participants distributed all coupons issued to them, not all coupons were redeemed and participants' knowledge about the redemption status was relatively accurate. For follow-up survey nonrespondents, it is possible that they knew that their recruitment effort was not successful either because they did not distribute coupons or the coupon recipient peer(s) did not participate, which further prompted them to be less motivated to participate in the follow-up.

6. Recruitment Intention and Behavior

Table 4 provides age and sex profiles of intended recruits reported in the main survey along with the profiles of intended recruits reported by those who were successful at recruitment and the profiles of actual recruits. In the PATH, intended recruits overall were younger than actual recruits: less than one out of five (16.8%) of the intended recruits were aged 60 or older, but almost half (41.8%) of the actual recruits were in that age category. When comparing intended recruits' age between successful recruiters and unsuccessful recruiters, there was a significant difference in age as the latter intended to recruit younger PWID ($\chi^2 = 36.9$; $df = 4$; $p < 0.001$). Sex was distributed similarly between intended

Table 3. Quasibinomial regression predicting of recruitment success.

	PATH				HLSK		
	Coeff	SE	*p* val.		Coeff	SE	*p* val.
Intercept	−0.573	0.482	0.236	Intercept	−0.760	0.267	0.005
Age (ref: 18–40 years)				Age (ref: 18–29 years)			
41–60 years	0.387	0.304	0.203	30–39 years	0.200	0.249	0.424
61+ years	0.623	0.343	0.070	40–49 years	−0.125	0.273	0.649
				50–59 years	−0.600	0.317	0.059
				60+ years	−0.540	0.404	0.182
Male vs. female	−0.244	0.200	0.223	Male versus female	−0.307	0.166	0.064
Non-Hispanic White versus other race	−0.336	0.293	0.252				
Living alone versus not education (ref: high school graduate)	0.005	0.226	0.983	Not married versus married	−0.488	0.219	0.026
				Education (ref: college degree)			
Less	−0.113	0.216	0.602	Less	0.158	0.196	0.418
More	−0.065	0.240	0.706	More	0.234	0.193	0.228
Employed versus unemployed	−0.118	0.298	0.691	Employed versus unemployed	−0.324	0.166	0.052
Income ≤ versus > USD20K	−0.277	0.308	0.369	Income ≤ versus > USD50K	0.079	0.169	0.642
				Interviewed in English versus Korean	−0.436	0.182	0.017
Site: Detroit versus other	−0.127	0.265	0.631	Site: Michigan versus Los Angeles	0.043	0.163	0.790
Network size[a]	0.007	0.006	0.288	Network size[b]	0.033	0.015	0.030
				Ethnic identity (ref: Korean)			
				Korean American	0.098	0.189	0.603
				Other	−0.043	0.396	0.914

[a]Number of PWID that participants know and interacted with more than once a week.

[b]Number of foreign-born Korean adults in Los Angeles/Michigan that participants know and feel close to.

Table 4. Profile of intended recruits and actual recruits.

	PATH			HLSK		
	Intended recruits of all recruiters % (n = 1074)	Intended recruits of successful recruiters % (n = 616)	Actual recruits % (n = 354)	Intended recruits of all recruiters % (n = 1138)	Intended recruits of successful recruiters % (n = 539)	Actual recruits % (n = 415)
Age						
18–29 years	13.3	10.7	8.2	32.6	30.4	34.1
30–39 years	20.0	16.1	11.0	25.2	26.4	25.4
40–49 years	16.0	14.9	13.8	20.8	21.0	21.0
50–59 years	33.8	38.2	25.1	15.4	16.1	14.0
60+ years	16.8	20.2	41.8	5.9	6.1	5.6
Gender						
Male	62.4	63.2	63.3	43.4	42.9	41.2
Female	36.4	35.9	36.7	56.6	57.1	58.8
Other[a]	1.2	1.0	-			

[a]Gay, Lesbian, Bisexual, Transgender, Something else.

and actual recruits and between successful recruiters' intended recruits and unsuccessful recruiters' intended recruits. In the HLSK, age and sex profiles were similar between intended and actual recruits, and there was no significant difference in intended recruits' profiles between successful and unsuccessful recruiters.

For follow-up participants, intended and reported recruits were similar with respect to age and sex consistently in the PATH and the HLSK (results not shown). Among follow-up participants who also successfully recruited, there was no difference across intended, reported and actual recruits in terms of these profiles in the HLSK; in the PATH, sex was similar, but age was different with the actual recruits being much older than intended as well as reported recruits, similar to Table 4. Overall, participants appeared to have distributed coupons to whom they intended. However, successful recruiters may target different types of peer(s). For example, age of the recipients mattered in the PATH as older recipients appeared to have participated at a higher rate than the counterparts.

7. Social Relationship within RDS Recruitment

The relationship between participants and their intended recruits and between participants and their actual recruiters is described in Table 5. PATH participants intended to recruit those that were not family members (94.2%), someone they felt close to (65.1%) and with whom (57.6%), and these patterns were not different between participants who were and were not successful at recruitment. This may make sense given that injection drug use was a key determinant of eligibility for the PATH. However, when examining participants' relationship with their recruiters, there was an increase in the close relationship, as well as in the co-injection compared to their relationship with intended recruits. This may mean that while PATH participants may have distributed or intended to distribute coupons to peers to whom they did not necessarily feel close to and did not inject drugs together, it was the coupon recipients who felt close to the recruiters or who did drugs with the recruiters that were more likely to participate than their counterparts. The fact that 93.6% of the PATH participants reported that they had contact with their recruiter last week may provide evidence of this.

With the HLSK, there was little to no difference in closeness between participants' relationship with intended recruits (also further divided by successful recruiters and unsuccessful recruiters) and with recruiters. However, the relationship type differed. About 28.0% of the intended recruits were family members, but it was 36.1% for successful recruiters and 20.5% for unsuccessful recruiters ($\chi^2 = 37.1$; $df = 2$; $p < 0.001$). Further, more than half (53.4%) of the participants reported that their recruiters were family members: 21.6% spouses, 7.7% parents, 7.9% siblings and 24.0% other family members (e.g., children). This means that recruitment coupons were redeemed at a higher rate when distributed to participants' families than to non-families.

Relationship types that follow-up participants had with their intended recruits were similar to those with reported recruits in both studies (results not shown).

Social relationship appears to matter in RDS recruitment, but may matter differently depending on the target group. For PWID, relationship closeness and substance co-injection mattered; but for Korean immigrants, it was the relationship type—non-family peers were recruited less successfully than family members.

Table 5. Participants' relationship with intended recruits and actual recruiters.

	PATH			HLSK		
	Intended recruits of all recruiters % (n = 1075)	Intended recruits of successful recruiters % (n = 616)	Recruiters % (n = 357)	Intended recruits of all recruiters % (n = 1138)	Intended recruits of successful recruiters % (n = 539)	Recruiters % (n = 415)
Relationship type						
Family	5.9	5.5	8.2	28.0	36.1	53.4
Spouse	1.0	0.8	1.7	10.6	15.3	21.6
Parent	0.4	0.5	0.3	7.0	11.0	7.7
Sibling	1.6	1.1	3.1	5.8	5.2	7.9
Other	2.9	3.1	3.1	4.6	4.7	24.0
Non-family	94.2	94.5	91.9	72.0	63.9	46.6
Friend	73.6	73.2	71.7	50.0	41.0	31.0
Romantic partner	1.8	1.0	1.1	4.0	4.7	3.8
Acquaintance	18.3	19.6	19.1	14.3	10.8	1.7
Colleague	0.5	0.7	-	7.7	7.5	6.3
Feel close to	65.1	67.1	81.2	88.8	92.3	87.0
Inject together	57.6	58.8	68.9	-	-	-
Contact within last week	-	-	93.6	-	-	78.4

8. Dynamics within Recruitment Process

Most follow-up participants reported that they distributed coupons within a week after issuance (96.2% for the PATH; 77.4% for the HLSK). In particular, 70.5% of the PATH participants distributed coupons within three days, and nearly all PATH participants reported distributing coupons in person. For the HLSK, the mode most frequently used for coupon distribution was Kakao talk, an extremely popular messaging app among Koreans, at 48.9%, followed by email (41.1%), text (24.2%), in-person (12.6%), phone conversation (4.3%), and others (1.7%), with 27.3% of the participants distributing coupons in more than one mode. WOM strategies used during recruitment in the PATH in order was: payment for participation (88.5%), survey topic (71.8%), payment for recruitment (62.2%), survey length (51.9%), study importance (20.5%) and something else (14.1%); and in the HLSK: payment (87.0%), Korean focus (72.6%), Web mode (50.9%), survey length (36.1%), study importance (20.9%) and something else (4.4%). Even though the recruitment reminders were sent to all HLSK participants whose issued coupons were not redeemed, 81.3% reported receiving the reminders. Among those who reported receiving reminders, 84.5% said they reminded their recruits about participation.

9. Implications for Fitness of RDS

One may wonder which of the two populations examined in this study fits better for RDS. Fitness, of course, should consider angles beyond the recruitment and recruitment processes that this study addresses. Assessing fitness of RDS for recruitment success requires subjecting populations to the same data collection protocols. Since data collection in our study followed different protocols (e.g., Web vs. in-person mode; a URL to access questionnaires vs. visiting an office; electronic versus paper coupons; up to two vs. three issued coupons), it is difficult to assess the fitness of RDS from the recruitment perspectives. Rather, our study offers important implications for RDS that may improve its general fitness, regardless of target populations. In particular, with coupon redemption rates around 30% commonly observed between two applications of RDS in this, as well as in other studies (e.g., Lee et al. 2017), at least three coupons need to be issued to each participant in order to prevent chains dying out. Otherwise, to meet operational goals, one may be forced to make unplanned protocol changes in the middle of data collection. However, issuing three coupons only guarantees no interruption in the overall recruitment *on average* and does not guarantee individual recruitment chains resembling Figure 1A. The systematic nature of recruitment noncooperation, discussed more in the next section, opens a door for integrating adaptive survey design to RDS. Creating recruitment chains like Figure 1A is difficult under current practices of RDS, where researchers have little to no control over the recruitment process. The adaptive survey design framework (Schouten et al. 2017) allows RDS operations to respond to the incoming data, such as survey data, coupon tracking data and paradata (e.g., interviewer observations). By capitalizing on the incoming data that allows predicting recruitment propensities at the individual level, researchers may change designs/protocols influential for producing chains like Figure 1A under the rules set prior to the data collection. This data-driven design adaptation approach may facilitate improved fitness of RDS across population types, so long as the target groups are networked.

10. Discussion

This study used data from two RDS studies (one targeting PWID in-person and the other targeting an ethnic minority group over the Web) and examined data collection progress, coupon use, profiles of intended versus actual recruits, social relationship between recruits and recruiters, and details of dynamics in the recruitment process. In both RDS studies, the coupon redemption rate was around 30%. Moreover, the recruitment success/cooperation differed systematically based on participants' characteristics. The lack of consistency in these patterns between the two RDS studies may suggest that the RDS recruitment cooperation is dependent on the target population and the context of survey administration.

Those who participated in the follow-up reported distributing most issued coupons to whom they intended. More often than not, participants knew their peers' coupon redemption status accurately. Interestingly, recruitment and recruitment success appeared to have been influenced by social relationship. The majority of recruiters were those with whom recruits had contacted within a week prior to their participation. For PWID, relationship closeness and whether using drugs together mattered for coupon recipients' participation. For Korean immigrants, coupons distributed to families were far more likely to be redeemed than those to non-families. Age and gender of intended recruits largely did not matter for the participation pattern, except for the study of PWID where older coupon recipients participated at a higher level than younger recipients.

Further, the timing of coupon distribution left little gap from the issuance, as most were reported to have been distributed within a week after their issuance in person for the in-person RDS and using a messaging app, email or text for the Web RDS. Incentive payment was the most prominent message participants told their peers during the recruitment process.

By no means does this study illustrate a complete picture of recruitment noncooperation in RDS. It uses information from the follow-up survey, which only about 40% of the participants completed. Albeit partial, the follow-up offers information about what happens between coupon issuance and redemption. For example, it provides answers to whether it is the RDS participants not distributing coupons or the coupon recipients not participating in the study that lead to recruitment noncooperation. Our study shows it is likely to be the latter potentially dictated by coupon recipients' perceived, as well as actual relationship with their coupon distributors. However, it is entirely possible that follow-up nonparticipants do not distribute coupons.

Taken together, this study suggests that "social networks" relevant to RDS recruitment may well be different than those discussed in the social network literature. In turn, this means that, without improving our ability to measure degrees specific to the chain-referral recruitment, weights used in RDS-specific estimators (e.g., Volz and Heckathorn 2008) are bound to be irrelevant and ineffective. Clearly, there are design features that participants highlight in their recruitment effort (e.g., incentives). Thorough investigations on participants' messaging and ways to leverage the design features affecting recruitment success also used in the messaging will allow us better design RDS studies. Moreover, as discussed with Figures 1A and 1B, two RDS studies of the same sample size starting from the same number of seeds do not mean the same recruitment chain structures. As done with

response rate calculations in survey research (American Association for Public Opinion Research 2016), needs for methodological transparency of RDS studies need to be recognized, and guidelines fostering such transparency need to be materialized (White et al. 2015).

11. Appendix

Table A1. Sample characteristics.

PATH % (n = 410)		HLSK % (n = 637*)	
Age		Age	
18–40 years	22.7	18–29 years	36.9
41–60 years	38.8	30–39 years	24.3
61+ years	38.5	40–49 years	20.9
		50–59 years	12.9
		60+ years	5.0
Gender		Gender	
Male	63.7	Male	39.4
Female	36.3	Female	60.6
Race/ethnicity			
Non-Hispanic White	29.8		
Other	70.2		
Living arrangement		Marital status	
Living alone	73.8	Married	54.6
Not living alone	26.2	Not married	45.4
Education		Education	
Less than high school	38.1	Less than college degree	30.0
High school graduate	35.1	College degree	40.0
More than high school	26.8	More than college degree	30.0
Employment status		Employment status	
Employed	11.5	Employed	55.4
Unemployed	88.5	Unemployed	44.6
Annual household income		Annual household income	
≤USD 20,000	90.1	≤USD 50,000	57.2
>USD 20,000	9.9	>USD 50,000	42.8
		Interview language	
		English	32.8
		Korean	67.2
Site		Site	
Detroit	69.5	Los Angeles	56.2
Macomb	4.6	Michigan	43.8
St. Clair	26.9		

* This excludes two participants whose survey responses were not properly stored in the data.

12. References

Abramovitz, D., E.M. Volz, S.A. Strathdee, T.L. Patterson, A. Vera, and S.D. Frost. 2009. "Using Respondent Driven Sampling in a hidden Population at Risk of HIV Infection: Who Do HIV-positive Recruiters Recruit?" *Sexually Transmitted Diseases* 36(12): 750–756. DOI: https://doi.org/10.1097/OLQ.0b013e3181b0f311.

American Association for Public Opinion Research. 2016. *Standard definitions: Final dispositions of case codes and outcome rates for surveys*. 9th edition. AAPOR. Available at: https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf (accessed September 2019).

Bostwick, W.B., T.L. Hughes, and B. Everett. 2015. "Health behavior, status, and outcomes among a community-based sample of lesbian and bisexual women." *LGBT Health* 2(2): 121–126. DOI: https://doi.org/10.1089/lgbt.2014.0074.

CDC. 2015. *National HIV Behavioral Surveillance: Injection Drug Use – Round 4 (NHBS-IDU4): Operations Manual*. Available at: https://www.cdc.gov/hiv/pdf/statistics/systems/nhbs/NHBS-IDU4-Operations-Manual-2015.pdf (accessed April 2018).

Gile, K.J. and M.S. Handcock. 2010. "Respondent-driven sampling: an assessment of current methodology." *Sociological Methodology* 40(1): 286–327. DOI: https://doi.org/10.1111/j.1467-9531.2010.01223.x.

Gile, K.J. and M.S. Handcock. 2015. "Network model-assisted inference from respondent-driven sampling data." *Journal of the Royal Statistical Society Series A, (Statistics in Society)* 178(3): 619–639. DOI: https://doi.org/10.1111/rssa.12091.

Gile, K.J., L.G. Johnston, and M.J. Salganik. 2015. "Diagnostics for respondent-driven sampling." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 178(1): 241–269. DOI: https://doi.org/10.1111/rssa.12059.

Groves, R.M. 1989. *Survey Errors and Survey Costs*. New York: John Wiley & Sons.

Groves, R.M. and M.P. Couper. 1998. *Nonresponse in household interview surveys*. New York: John Wiley & Sons.

Hafeez, S. 2012. *A review of the proposed STROBE-RDS reporting checklist as an effective tool for assessing the reporting quality of RDS studies from the developing world*. London, UK: LSHTM.

Hathaway, A.D., E. Hyshka, P.G. Erickson, M. Asbridge, S. Brochu, M.M. Cousineau, C. Duff, and D. Marsh. 2010. "Whither RDS? An investigation of respondent driven sampling as a method of recruiting mainstream marijuana users." *Harm Reduction Journal* 7(1): 15. DOI: https://doi.org/10.1186/1477-7517-7-15.

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

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.

Kalsbeek, W.D. 2003. "Sampling minority groups in health surveys." *Statistics in Medicine* 22: 1527–1549.

Kalton, G. and D.W. Anderson. 1986. "Sampling rare populations." *Journal of Royal Statistical Society, Series A* 149(1): 65–82. DOI: http://dx.doi.org/10.2307/2981886.

Kalton, G. and I. Flores-Cervantes. 2003. "Weighting methods." *Journal of Official Statistics* 19(2): 81–97. Available at: https://www.scb.se/contentassets/ca21efb41-fee47d293bbee5bf7be7fb3/weighting-methods.pdf (accessed February 2020).

Lee, S. 2009. "Understanding respondent driven sampling from a total survey error perspective." *Survey Practice* 2(6) 1–6. DOI: https://doi.org/10.29115/SP-2009-0029.

Lee, S., A.R. Ong, and M. Elliott. 2018. "Two applications of respondent driven sampling: Ethnic minorities and illicit substance users." Paper presented at the Workshop on Improving Health Research for Small Populations. National Academy of Sciences, Engineering and Medicine, Washington, DC, U.S.A. January 2018. Available at: http://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_185285.pdf (accessed September 2019).

Lee, S., Z.T. Suzer-Gurtekin, J. Wagner, and R. Valliant. 2017. "Total survey error and respondent driven sampling: Focus on nonresponse and measurement errors in the recruitment process and the network size reports and implications for inferences." *Journal of Official Statistics* 33(2): 335–366. DOI: https://doi.org/10.1515/jos-2017-0017.

Lee, S., Z.T. Suzer-Gurtekin, J. Wagner, and R. Valliant. 2012. "Exploring error properties of respondent driven sampling." Paper presented at the Joint Statistical Meeting, San Diego, CA, U.S.A. July 2012.

Lee, S., J. Wagner, R. Valliant, and S. Heeringa. 2014. "Recent developments of sampling hard-to-reach populations: an assessment." In *Hard to Survey Populations*, edited by R. Tourangeau, B. Edwards, T. Johnson, and K. Wolter: 424–444. Cambridge, UK: Cambridge University Press.

Li, J., T.W. Valente, H.S. Shin, M. Weeks, A. Zelenev, G. Moothi, H. Mosher, R. Heimer, E. Robles, G. Palmer, and C. Obidoa. 2017. "Overlooked threats to respondent driven sampling estimators: peer recruitment reality, degree measures, and random selection assumption." *AIDS and Behavior* 22(7): 2340–2359. DOI: https://doi.org/10.1007/s10461-017-1827-1.

Lu, X., L. Bengtsson, T. Britton, M. Camitz, B.J. Kim, A. Thorson, and F. Liljeros. 2012. "The sensitivity of respondent-driven sampling." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 175(1): 1–26. DOI: https://doi.org/10.1111/j.1467-985X.2011.00711.x.

Luthra, R. 2011. "*RDS for Migration Studies? A Review and Invitation to Discuss*." Paper presented at the Workshop on Design, Implementation, and Analysis: An Exploration of Respondent Driven Sampling, London, UK.

Martin, K., T.P. Johnson, and T.L. Hughes. 2015. "Using respondent driven sampling to recruit sexual minority women." *Survey Practice* 8(2): 273. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5066809/ (accessed April 2020).

McCreesh, N., A. Copas, J. Seeley, L.G. Johnston, P. Sonnenberg, R.J. Hayes, S.D.W. Frost, and R.G. White. 2013. "Respondent driven sampling: determinants of recruitment and a method to improve point estimation." *PLoS ONE* 8(10): e78402. DOI: https://doi.org/10.1371/journal.pone.0078402.

Selvaraj, V., K. Boopathi, P. Paranjape, and S. Mehendale. 2016. "A single weighting approach to analyze respondent-driven sampling data." *The Indian Journal of Medical Research* 144(3): 447–459. DOI: https://doi.org/10.4103/0971-5916.198665.

Schouten, B., A. Peytchev, and J. Wagner. 2017. *Adaptive Survey Design*. Boca Raton, FL: CRC Press.

Shadish, W.R. and T.D. Cook. 1999. "Design rules: More steps towards a complete theory of quasi-experimentation." *Statistical Science* 294–300.

Singer, E. 2002. "The use of incentives to reduce nonresponse in household surveys." In *Survey Nonresponse,* edited by R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little, 163–178. New York, NY: Wiley. 163–177

Stein, M.L., V. Buskens, P.G.M. van der Heijden, J.E. van Steenbergen, A. Wong, M.C.J. Bootsma, and M.E.E. Kretzschmar. 2018. "A stochastic simulation model to study respondent-driven recruitment." *PLoS One* 13(11): e0207507. DOI: https://doi.org/10.1371/journal.pone.0207507.

Strömdahl, S., X. Lu, L. Bengtsson, F. Liljeros, and A. Thorson. 2015. "Implementation of Web-based respondent driven sampling among men who have sex with men in Sweden." *PLoS ONE* 10(10): e0138599. DOI: https://doi.org/10.1371/journal.pone.0138599.

Tomas, A. and K.J. Gile. 2011. "The effect of differential recruitment, non-response and non-recruitment on estimators for respondent-driven sampling." *Electronic Journal of Statistics* 5: 899–934. DOI: https://doi.org/10.1214/11-EJS630.

Tucker, C., M.P. Cohen, A. KewalRamani, and S. Eyster. 2015. "Surveying the District of Columbia GLBT community using respondent-driven sampling." Paper presented at the annual meeting of the American Association for Public Opinion Research, May 2015. Available at: http://www.aapor.org/AAPOR_Main/media/AnnualMeetingProceedings/2015/A3-3-Tucker.pdf (accessed September 2019).

Verdery, A.M., M.G. Merli, J. Moody, J. Smith, and J.C. Fisher. 2015. "Respondent-driven sampling estimators under real and theoretical recruitment conditions of female sex workers in China." *Epidemiology* 26(5): 661–665. DOI: https://doi.org/10.1097/EDE.0000000000000335.

Volz, E. and D.D. Heckathorn. 2008. "Probability based estimation theory for respondent driven sampling." *Journal of Official Statistics* 24(1): 79–97. DOI: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/probability-based-estimation-theory-for-respondent-driven-sampling.pdf (accessed May 2020).

Wagner, J. and S. Lee. 2014. "Sampling rare populations." In *Handbook of Health Survey Methods*, edited by T.P. Johnson, 77–104. Hoboken, N.J.: Wiley.

White, R.G., A.J. Hakim, M.J. Salganik, M.W. Spiller, L.G. Johnston, L. Kerr, C. Kendall, A. Drake, D. Wilson, K. Orroth, M. Egger, and W. Hladik. 2015. "Strengthening the reporting of observational studies in epidemiology for respondent-driven sampling studies: "STROBE-RDS" statement." *Journal of Clinical Epidemiology* 68(12): 1463–1471. DOI: https://doi.org/10.1016/j.jclinepi.2015.04.002.

World Health Organization and UNAIDS. 2013. *Introduction to HIV/AIDS and sexually transmitted infection surveillance: Module 4: Introduction to Respondent Driven Sampling*. Geneva, Switzerland. Available at: http://applications.emro.who.int/dsaf/EMRPUB_2013_EN_1539.pdf (accessed June 2017).

Received November 2018
Revised April 2019
Accepted November 2019

sciendo

Journal of Official Statistics, Vol. 36, No. 2, 2020, pp. 361–378, http://dx.doi.org/10.2478/JOS-2020-0019

Measuring the Sustainable Development Goal Indicators: An Unprecedented Statistical Challenge

Steve MacFeely[1]

In March 2017, the United Nations (UN) Statistical Commission adopted a measurement framework for the UN Agenda 2030 for Sustainable Development, comprising of 232 indicators designed to measure the 17 Sustainable Development Goals (SDGs) and their respective 169 targets. The scope of this measurement framework is so ambitious it led Mogens Lykketoft, President of the seventieth session of the UN General Assembly, to describe it as an 'unprecedented statistical challenge'.

Naturally, with a programme of this magnitude, there will be foreseen and unforeseen challenges and consequences. This article outlines some of the key differences between the Millennium Development Goals and the SDGs, before detailing some of the measurement challenges involved in compiling the SDG indicators, and examines some of the unanticipated consequences arising from the mechanisms put in place to measure progress from a broad political economy perspective.

Key words: 2030 Agenda; unintended consequences; national statistical systems; administrative data.

1. Introduction

The 2030 Agenda for Sustainable Development and the Sustainable Development Goals (SDG) Global Indicator Framework (GIF) was adopted by the United Nations (UN) Statistical Commission in March 2017 (UN Statistical Commission 2017) and subsequently adopted by the UN General Assembly in July 2017 (UN General Assembly 2017). The framework comprises 232 statistical indicators designed to measure the seventeen 2030 Agenda goals and their respective 169 targets. The aim of the GIF is to provide good quality, verifiable evidence on progress towards achieving the 2030 Agenda. However, populating those indicators and providing that evidence poses enormous challenges. So much so, it led Mogens Lykketoft, President of the seventieth session of the UN General Assembly, to describe it as an 'unprecedented statistical challenge' (Lebada 2016).

This article outlines some of the measurement challenges involved in compiling the SDG indicators, some of the tensions that have arisen during the process to date, and also examines some of the unanticipated consequences arising from the mechanisms put in place to measure progress. By explaining, very briefly, some of the key differences between the Millennium Development Goals (MDGs) and the SDGs, and the political circumstances

[1] United Nations Conference on Trade and Development, Palais des Nations, CH-1211 Geneva 10, Switzerland. Email: Steve.macfeely@un.org

in which the GIF came to life, the article attempts to explain why this project is indeed an unprecedented statistical challenge.

The remainder of this article is presented in nine sections. The next section identifies some of the most important differences between the SDGs and their predecessor, the MDGs. The following two sections outline some of the challenges and tensions that have emerged in measuring the SDGs. Sections 5 and 6 explain how the Inter-agency and Expert Group on SDG Indicators (IAEG-SDG) has classified the SDG indicators into tiers, and details also some of the cost estimates for compiling SDG indicators put forward by several commentators. Section 7 outlines some of the unintended consequences that are emerging from the process. The penultimate section presents a discussion of some issues arising, before the article is concluded with some recommendations.

2. From MDGs to SDGs

At the beginning of 2016, the UN SDGs replaced the MDGs, which had been in place since the turn of the century. Although both sets of goals describe an aspirational road map for global development, the SDGs came about through a profoundly different process than the MDGs, which was essentially a distillation of the major agreements from the main development conferences of the 1990s, (such as the World Summit for Children 1990; the UN Conference on Environment and Development 1992; the World Conference on Human Rights 1993; the International Conference on Population and Development 1994; the World Summit for Social Development 1995; the Fourth World Conference on Women 1995; the Second UN Conference on Human Settlements Habitat II 1996; the World Food Summit 1996). These agreements were compiled by the UN Secretariat and reflected in the UN Secretary General's Millennium Report, *We the Peoples: The Role of the United Nations in the 21st Century* (Annan 2000), which outlined the challenges for development in a globalised world. At the fifty-fifth General Assembly, designated the 'Millennium Summit', 189 Member States adopted the Millennium Declaration (UN 2000). This Declaration committed nations to reduce extreme poverty by 2015. The following year, in August 2001, the UN Secretariat published the final set of eight MDGs. It was described by Hulme (2009, 4) as 'the world's biggest promise.'

The SDGs, by contrast, emerged from the shadow of the MDGs, which had been criticised for pushing a donor-driven agenda, excluding any discourse critical of the Washington Consensus and not fully reflecting the will or views of peoples or governments. From the outset, the SDG process aimed to create a people-centred development agenda. To do so, an unprecedented global consultation was undertaken. Following three years of consultation and negotiation, involving thousands of people, *Transforming Our World: The 2030 Agenda for Sustainable Development* (UN 2015a) was formally adopted by 193 heads of government, including 150 heads of state on 25 September 2015. The 2030 Agenda adopted a broad view of development, one that encompassed not just ending extreme poverty and eradicating hunger, but one that aspires to foster global prosperity in an economically and environmentally sustainable and equitable way. The 17 SDGs and their 169 targets would be 'action oriented, global in nature and universally applicable' (UN, 2013, 4), and were described by Ban Ki-moon (UN 2015b), former Secretary General of the UN, as the 'to do list for planet and people'.

3. Some Challenges in Measuring the SDGs

From a statistical perspective, the implications of the 2030 Agenda and the accompanying GIF are enormous. Not only have the number of goals and targets increased considerably compared with the MDGs (The MDGs had 8 goals, 21 targets and 60 indicators, whereas the SDGs have 17 goals, 169 targets and 232 indicators), but so also has the complexity of these targets. The scope of the 2030 Agenda is also far wider than that of its predecessor, attempting to span the full spectrum of development issues, including not only aspects of society, economy and the environment, but also institutional coordination.

A simple illustration of the complexity is available from the report *Data Disaggregation and SDG Indicators: Policy Priorities and Current and Future Disaggregation Plans* (IAEG-SDG, 2019). This matrix details the minimum set of disaggregation required for each indicator. The level of disaggregation varies considerably by indicator. For example, the "minimum required disaggregation dimension" demanded by Target 1.3 (Implement nationally appropriate social protection systems and measures for all, including floors, and by 2030 achieve substantial coverage of the poor and the vulnerable) is: sex; age; employment status; disability status; pregnancy; work-injury victims; and income. For Target 10.2 (By 2030, empower and promote the social, economic and political inclusion of all, irrespective of age, sex, disability, race, ethnicity, origin, religion or economic or other status) the "minimum required disaggregation dimension" is: sex; age; disability status; race; ethnicity; origin; religion; and other economic or social status.

The first challenge facing statisticians was to clarify what it was they were being asked to measure. This was easier said than done. Deciphering or interpreting exactly what is meant by the agreed text of *Transforming our World: The 2030 Agenda for Sustainable Development* (UN 2015a) was not always straightforward. Lack of clear definitions and inconsistent use of terminology are just some examples of where statisticians, in selecting appropriate indicators, were forced to decide what the targets actually meant. For example, what is meant by 'sustainable'?

Does it just mean environmentally sustainable, or does it also mean economically sustainable, or socially sustainable? Environmentalists will naturally assume it means environmentally sustainable, but economists will equally assume it means economic sustainability. The next question is how long a trend should be exhibited before it can be considered sustainable – will this be the same for economic or environmental variables? What about 'economic stability'? Target 17.13 calls for global macro-economic stability. Although there is no consensus on what this means, it has been agreed it will be measured by a dashboard of indicators. The composition of this dashboard will effectively determine whether the 2030 Agenda adopts an orthodox or heterodox view of the global economy.

What are the 'basic services' or the 'new technologies' referred to in Target 1.4 (By 2030, ensure that all men and women, in particular the poor and the vulnerable, have equal rights to economic resources, as well as access to basic services, ownership and control over land and other forms of property, inheritance, natural resources, appropriate new technology and financial services, including microfinance) and are they the same in all parts of the world? This might seem like pedantry, but it matters when you are trying to design an appropriate measurement. A plethora of seemingly commonly understood words, such as: access; adverse; adequate; appropriate; basic; benefit; efficient; effective;

informal; infrastructure; integration; promote; resilience; resource; sustainable; and vulnerable caused comprehension problems and challenges of consistent interpretation across the 169 targets, requiring the construction of a SDG ontology (UN Environmental Programme 2015) to make progress.

Another challenge is the lack of priority within complex and sometimes rather muddled targets. This has proven particularly thorny, as statisticians were instructed by their political masters to limit the number of indicators to one indicator per target. Numerate readers will have noted that this guideline was not respected, as 169 targets resulted in 232 indicators. In truth, to measure the targets properly, closer to 500 indicators would probably be required. Take Target 17.19 (By 2030, build on existing initiatives to develop measurements of progress on sustainable development that complement GDP and support statistical capacity building in developing countries) for example. This target combines two completely different and unrelated issues: firstly, the measurement of progress beyond GDP and secondly, supporting statistical capacity-building. This bundling, not uncommon to many targets, poses a dilemma. Which element of the target should be measured? Both are very important, but both are also very complex. The challenge of how to properly measure progress is a highly contentious issue, hotly debated by economists, social scientists, environmentalists and statisticians (MacFeely 2016), and would probably need a whole dashboard of indicators to do justice to this one issue. Equally, the best way to approach statistical capacity-building is also being actively discussed and reassessed (Jütting 2016). The idea that such a cocktail of issues could sensibly be amalgamated into a single indicator is absurd. The Economist (2015b), citing Target 4.7 (By 2030, ensure all learners acquire the knowledge and skills needed to promote sustainable development, including among others through education for sustainable development and sustainable lifestyles, human rights, gender equality, promotion of a culture of peace and non-violence, global citizenship and appreciation of cultural diversity and of culture's contribution to sustainable development) as an example, put it bluntly, simply saying, 'try measuring that.'

Unsurprisingly, indicator 4.7.1 (Extent to which (i) global citizenship education and (ii) education for sustainable development, including gender equality and human rights, are mainstreamed at all levels in: (a) national education policies; (b) curricula; (c) teacher education; and (d) student assessment) has been classified as Tier 3 by the IAEG-SDGs (see Section 5).

Although the scope of the 2030 Agenda is universal and applies to all countries, clearly not all targets are relevant to every country. Striking a balance between national and global demands has proven challenging. For example, Target 3.3 (By 2030, end the epidemics of AIDS, tuberculosis, malaria and neglected tropical diseases and combat hepatitis, water-borne diseases and other communicable diseases) targets the eradication of a wide variety of diseases, many of which are not prevalent across the globe. As a result, statisticians have selected two statistical indicators, targeting HIV and tuberculosis, as the appropriate global indicators. So not all elements of the target are addressed and thus some elements of the target must be ignored and remain unquantified. While this might make sense from a global perspective, it may not necessarily make sense from a regional or national viewpoint. For example, the control of dengue fever is not a big issue globally, but is very important in South-East Asia. Not surprisingly, when the dust settled, and the indicators had been selected, researchers criticised the indicators for being reductionist (Mair et al. 2018).

Other important decisions are still to be taken. For example, how will changes in the composition of groups be dealt with. Over the course of fifteen years, several Least Developed Countries (LDCs) are likely to graduate from that status. According to the UN Department of Economic and Social Affairs (2019), Vanuatu is expected to graduate in 2020, Angola in 2021, Bhutan in 2023 and São Tomé and Príncipe, the Solomon Islands and perhaps Bangladesh in 2024. What are the implications of this for time-series analyses? Twenty-four of the 169 targets explicitly mention LDCs. When we target an annual growth in GDP of 7% in the LDCs (Target 8.1), a doubling in the share of employment in industry for LDCs (Target 9.2), or a doubling of LDCs' share of global exports (Target 17.11), which LDC group are we referring to? Will rates of change be calculated using the original composition in 2015 or the group as it will be composed in 2030? Or will both series be presented side by side? A relatively straightforward decision, but one where the choice will, most likely, lead to quite different results and may open considerable room for the interpretation of success.

4. Some Tensions in Measuring the SDGs

A surprising discovery emerged during the preparatory work to develop the SDG GIF; many Member States appeared not to fully understand the distinction between national and international official statistics and the significance or purpose of having both. This misunderstanding extended beyond political circles and included also representatives from national statistical offices (NSOs). The SDGs brought this distinction into sharp focus. Confusion around this issue, and subsequent tensions became most acute during the discussions on formulating the '*Guidelines on Data Flows and Global Data Reporting for Sustainable Development Goals*' document (IAEG-SDG 2018b).

Arguably, many of these tensions could have been avoided, if early in the process, the exact scope and purpose of the SDG GIF had been communicated clearly to Member States, and the distinction between global and national indicators had been made clear. Kapto (2019, 135) summarised it well, saying "A tense debate is taking place on data flows from national to regional to global levels, and on custodian agencies' role in harmonising national data for global comparability, as countries assert their sovereignty over national data." The insistence by some Member States that official country data should be prioritised may ultimately be counter-productive given the paucity of data available in many developing countries, resulting in many SDG indicators remaining unpopulated. It may also, inadvertently, undermine the role of international organisations (IOs) that play an important role in compiling harmonised official international statistics, which often involves amending or imputing national data. The 'country first' approach, while to some extent understandable, is nevertheless somewhat incongruous with statements the same countries make vis-à-vis the importance of harnessing the data revolution or using big data and geo-spatial information.

Nevertheless, countries anxious to keep control over messaging are determined that only official national data are used to populate the SDG indicators. Apart from communication control, there are, of course, some legitimate reasons why: (1) national data may be superior, from the perspective of policy formation, as they can be integrated with other national data to present a coherent story; (2) as already noted, the SDG process

has expanded the frontiers of official statistics and NSOs may wish to retain control so that they can develop expertise in new statistical domains; (3) NSOs get frustrated when they find results they don't recognise in international databases – often the reasons for the differences are legitimate, but have not been communicated to countries (of course, in some cases, the results were communicated, but some countries didn't pay attention).

There are, however, some circumstances where the 'country first' approach may not necessarily be the best approach. Targets, such as 16.5 (Substantially reduce corruption and bribery in all their forms) or 16.6 (Develop effective, accountable and transparent institutions at all levels), which deal with corruption, bribery and the accountability of institutions, provide perfect examples of why it might make sense to use external or unofficial data, as official data may not exist or may not be sufficiently trustworthy to provide an independent, impartial picture of such sensitive matters. Another exception might be where a single source could provide better-quality and globally more consistent data than the amalgamation of multiple individual country data sets. This might be applicable to targets such as 15.1 that deal with forest, drylands, wetlands and mountain regions governed by international agreements. Arguably, superior quality and internationally comparable data could be derived from satellite imagery.

Using alternative sources to compile official national statistics might also be reasonable where problems with data exist. Problems with data could mean anything from errors or inaccuracies, non-adherence to international standards, incompleteness or data gaps, inconsistencies over time, or imbalances. A good example of where this might arise is the asymmetries that frequently exist between bilateral trade data sets. From a global perspective, unbalanced trade data are not especially useful, and so steps are taken to remove these asymmetries. However, this may lead to a mismatch between official national statistics and official international statistics. This issue is not unique to international trade, 'problems' with national data exist across a range of statistical domains. For the moment, the challenge of how to balance the needs of national and global interests remains unresolved. However, the discussion should not be characterised as national versus international official statistics, but rather how best to integrate and use different statistics to deliver on requirements.

Despite the best efforts of NSOs and IOs, internationally comparable data remains a challenge. The SDG process has exacerbated this challenge, as many of the targets, and consequent indicators, fall well outside the scope of traditional official statistics and thus, are not guided by agreed international measurement standards. Even for those indicators that do fall within the scope of traditional official statistics, there will be a wide variety in general quality and adherence to international standards across countries.

The goals and targets of the 2030 Agenda are underpinned by the ambition that 'no one gets left behind' (UN 2015a). This ambition was translated for statisticians by Mogens Lykketoft, President of the seventieth session of the UN General Assembly, as 'leaving no one uncounted' (Lebada 2016). In principle, this is fine, but such a literal translation does not make much sense from a statistical perspective. The purpose of official statistics, with a few exceptions, such as population censuses, is not to account for every single person but rather to provide general aggregate, anonymised information on population cohorts of interest. This is a fundamental difference between producing official statistics and audited accounts. Apart from issues of confidentiality, the cost of realising the ambition of 'leaving no one uncounted' would be prohibitive and not financially viable for even the best-resourced and

most efficient statistical systems. The challenge for the global statistical system is how to sufficiently improve the granularity of data in a way that prioritises the measurement of the poorest and most vulnerable, that does not divert scarce resources into generating fruitless levels of disaggregation and yet satisfies the demands of political rhetoric.

5. Classifying the SDG Indicators

The far-reaching ambitions of the 2030 Agenda have led to development targets that are well ahead of available official statistics and statistical concepts. In many cases, appropriate statistical definitions and methodologies do not yet exist from which to generate indicators. To elaborate this problem and facilitate the population of the GIF, the IAEG-SDG (2018a) has classified all SDG indicators into three tiers based on their conceptual development and availability of data. The tiers are:

> Tier 1: the indicator is conceptually clear, has an internationally established methodology, standards are available, and data are regularly produced by countries for at least 50% of countries and of the population in every region where the indicator is relevant.
>
> Tier 2: the indicator is conceptually clear, has an internationally established methodology, standards are available, but data are not regularly produced by countries.
>
> Tier 3: no internationally established methodology or standards are yet available for the indicator, but methodology/standards are being (or will be) developed or tested.

In September 2019, the IAEG-SDG reported that 45% of the selected indicators were classified as Tier 1 (see Table 1). Furthermore, they reported that 14% of the indicators remained classified as Tier 3. While Table 1 shows the not inconsiderable improvements in conceptual development and data availability that has been made since 2016, it also highlights the magnitude of the task still facing the global statistical community. The pace of transition of indicators through the tiers to reach Tier 1 is likely to slow, as presumably the low hanging fruit will be picked first. Table 1 suggests this is indeed the case, as the conversion rate to Tier 1 was slower between December 2017–2018 than between December 2016–2017. Between December 2018 and September 2019 only three indicators were converted to Tier 1. A further cautionary footnote should be added. Research undertaken by Dang and Serrjuddin (2019) of the World Bank

Table 1. Number of SDG indicators by tier.

Tier Classification	December 2016		December 2017		December 2018		September 2019	
	Number	%	Number	%	Number	%	Number	%
1	81	35	93	40	101	44	104	45
2	57	25	66	28	84	36	89	38
3	88	38	68	29	41	18	33	14
Multiple	4	2	5	2	6	3	6	3
Total	230	100	232	100	232	100	232	100

Source: Derived from IAEG-SDG (2019b). https://unstats.un.org/sdgs/iaeg-sdgs/tier-classification/

highlights the 'overwhelming challenge with missing data' and suggests that not all Tier 1 indicators are actually populated. They estimate that only 19% of the required GIF data are currently available.

6. The Cost of Measurement

One of the implications of such a broad and ambitious development agenda is the price tag. Estimates vary, but Ambassador Macharia Kamau of Kenya, who co-chaired the SDG intergovernmental consultative process, estimates that implementing the SDG agenda could cost somewhere between USD 3.5 trillion and USD 5 trillion per year (Deen 2015). The Economist (2015a) described their estimate, of between USD 2 trillion and USD 3 trillion per year (or the equivalent of 4% of global GDP), as 'unfeasibly expensive'. The Intergovernmental Committee of Experts on Sustainable Development Financing (2014) estimated the value of investment in infrastructure required to achieve the eradication of poverty alone at between USD 5 trillion and USD 7 trillion annually.

Even for developed countries with relatively advanced and sophisticated statistical systems, the demands arising from the SDG monitoring framework are immense. When one considers that in 2019, only 45% of the proposed 232 indicators were classified as Tier 1 (see Table 1), the extent of the problem becomes clear. PARIS21 (2015, 11) has estimated that 'funding for statistics needs to be increased from current commitments of between USD 300 million and USD 500 million to between USD 1 billion and USD 1.25 billion by 2020'. The Global Partnership for Sustainable Development Data estimates that around USD 650 million per year is needed to collect data, of which only USD 250 million is currently funded (Runde 2017). Irrespective of which estimate is used, these sums clearly exceed existing funding (UNCTAD 2016). While clearly the bulk of these resources will be required to improve statistical capacity in developing countries, it is evident that resources will be required in the developed world too in order to deliver on the promises made by national governments.

Although statistics account for only 0.3% of official development assistance (ODA) (PARIS21 2017), USD 541 million is not a trivial amount of money and the thought of paying more doesn't appear to excite many donors. In this context, Slotin (2018) asks a relevant question 'if development data is so powerful, why does no one want to pay for it?' Assuming the answer is a poor understanding of the contribution of official statistics as a public good to democracy, commerce and social wellbeing, official statisticians have set out to show it's a price worth paying, arguing 'if you think statistics are expensive – try ignorance'. To try to justify this claim and the costs of measurement, a nascent industry is now emerging, where statisticians are trying to estimate the benefits of official statistics. Chui et al. (2013) found that open data globally could potentially unlock between USD 3.2 trillion and USD 5.4 trillion in economic value per year. UNECE (2019) has compiled a report on the various methods used and cites a variety of estimates. The numbers are seductive, but it doesn't change the fact that, as yet, NSOs are not getting additional funding to compile SDG indicators.

7. Some Unanticipated Consequences

According to the American cultural and intellectual historian T.J. Jackson Lears, 'All history is the law of unintended consequences' (Cohen 2013). It should not be surprising

then that a development plan as broad as the 2030 Agenda and the implementation of the SDG GIF should throw up a few surprises. This section of the article examines what some of these surprises are and what the consequences might be.

The delegation of the selection and measurement of the statistical indicators to the UN Statistical Commission was a major triumph for official statistics. It was an explicit recognition of the need for apolitical, independent and impartial official statistics to measure progress and the separation of function between statistical compilers and statistical consumers. It also responded to the views expressed by many that the 2030 Agenda needed an effective performance system with clear metrics measuring progress towards each goal (Warren 2015; Costanza et al. 2016; Jacob 2017).

However, SDG targets are not 'targets' in the normal sense of the word – they are, for the most part, not clear time-delimited objectives but rather general, often complex, aspirations that leave generous space for interpretation. Furthermore, they incorporate all the unresolved issues left over from the negotiation phase. Thus, in handing over the measurement task to the statistical community, the interpretation of the targets was effectively delegated too. Many heads of state and policy mandarins might be surprised by, or even contest, this statement. However, it was statisticians who selected the indicators that specifically defined what the 2030 Agenda text actually meant. This is an important point because the SDG indicators do not simply measure the 2030 Agenda, they define it. As noted above, the composition of the dashboard selected to measure Target 17.13 will effectively determine whether the 2030 Agenda adopts an orthodox or heterodox view of the global economy. Equally, how statisticians interpret the word 'illicit' when designing indicator Target 16.4 (Total value of inward and outward illicit financial flows in current USD) will determine whether corporate profit shifting is included or not. In making these decisions, statisticians are effectively determining what the SDG targets mean. The indicators selected are the performance metrics for the 2030 Agenda and thus will have direct consequences for whether the 2030 Agenda is judged a success or a failure.

With 169 SDG targets, many of which are multidimensional, there was an understandable fear of indicator proliferation. After all, the MDGs had only 21 targets but 60 indicators, a ratio of 3:1. The prospect of 500 plus indicators was not attractive to politicians. Hence the limit of one indicator per target. Yet there are consequences to measuring a multifaceted target with a single indicator. The first and most obvious being that, unless a composite indicator or a multidimensional dashboard can be designed, several elements of the target will be sacrificed. This may be appropriate if the target dimensions are all somehow related. However, in several cases, the dimensions included in the targets do not appear to be related at all, in which case, any single indicator will be problematic. A single indicator thus introduces the risk that unmeasured aspects of a target will be ignored, and interconnections between different elements of the target (or other targets) will remain unquantified. An obvious worry from a policy perspective are the implications for consequent behaviour and the risk that *only* what gets measured gets done.

Another concern is how to extrapolate from narrow indicators to broad targets. For example, Target 17.4 (Assist developing countries in attaining long-term debt sustainability through coordinated policies aimed at fostering debt financing, debt relief and debt restructuring, as appropriate, and address the external debt of highly indebted

poor countries (HIPCs) to reduce debt distress) is a complex, multidimensional target represented by a single indicator. In this case, a variety of complex issues, such as: long-term debt sustainability; debt financing; debt relief; debt restructuring; and external debt have all been shoe-horned into a single indicator. Furthermore, indicator 17.4.1 doesn't really address any of the target elements directly or adequately, raising questions as to how progress towards Target 17.4 should be interpreted.

Very few of the SDG indicators are bespoke indicators that fit the specifications of the target exactly. Furthermore, very few were deliberately designed for the purpose for which they are now being used; most are to some extent or other, recycled, proxies. This will matter when the trends and patterns identified by the indicator are extrapolated and applied to all elements of the broader target. It is important to understand what that original purpose of the indicator was, so that its appropriateness as an SDG indicator can be assessed. The small print (otherwise known as metadata) will be very important when analysing the SDG results.

8. Discussion

The 2030 Agenda may have a profound influence on the shape and organisation of official statistics in the future. As noted above, many policy discussions are running far ahead of available statistics, and so the SDGs are likely to be the driving force, or *raison d'être*, for many statistical advances in the coming years, both in terms of statistical concepts and methodology but also in terms of statistical organisation and the use of new data sources. It is very important that all national statistical systems (NSSs) engage actively in these discussions. As Harari (2018, ix) notes 'history gives no discounts.' Countries that do not engage will not be exempt from the consequences. This may have three unexpected outcomes:

1. Statistical organisation – the demand for new statistics may inadvertently open the door to the outsourcing or privatisation of official statistics if the existing system fails to deliver on the huge expectations that appear to exist. There is a risk that if the UN statistical system cannot fill the vacuum created by the Tier II and III indicators, then someone else will. This is not necessarily a bad thing; there are many who argue that a more inclusive approach (MacFeely and Nastav 2019) or incorporating citizen science (Fritz et al. 2019) might benefit the SDG measurement process.
2. Big Data – there is growing pressure on NSOs to try to harness big data to compile statistics. This is, in of itself not problematic, although expectations should be realistic (see MacFeely 2019), but it may distract from developing administrative data sources, which arguably will be more useful (in the short to medium term at least). The real challenge, and optimal objective, is for NSOs to find a way to integrate multiple data sources, whether traditional or new, to develop efficient national statistical and information systems (UNECE 2016; MacFeely and Barnat 2017).
3. Reputational risk – given the very short timeframe in which the GIF was developed without any appreciable additional resources, a lot has been achieved. Nevertheless, the SDG GIF may still disappoint the high expectations, and this in turn may undermine the UN Statistical Commission (MacFeely and Nastav 2019). On the

other hand, it also offers an opportunity to re-engage with policy makers and discuss the importance of official statistics.

Thus, for a variety of reasons the SDG GIF is likely to have lasting implications beyond the 2030 Agenda. NSOs need to reflect carefully on these issues. A particular challenge posed by the 2030 Agenda for statisticians is that some of the SDG targets deal with phenomena that arguably cannot be measured comprehensively, if at all. Cited above, Target 4.7 is a good example. Target 17.16 (Enhance the Global Partnership for Sustainable Development, complemented by multi-stakeholder partnerships that mobilize and share knowledge, expertise, technology and financial resources, to support the achievement of the Sustainable Development Goals in all countries, in particular developing countries) is another. This is not a criticism of the target or the aspirations contained therein, simply that some issues are by their nature nebulous and defy robust quantitative measurement. As quantification and metrics have irrevocably become part of society's zeitgeist, no one is questioning whether this approach is sensible or achievable – it is now a commonly held view that everything can and must be measured. In an era of governance by numbers, the management clichés of 'measure what you treasure' or 'what gets measured gets done' rule supreme. However, as Muller (2018, 8) points out, 'measurement may become counterproductive when it tries to measure the unmeasurable and quantify the unquantifiable.' While no one can credibly challenge the logic of evidence-informed decision making, arguably statisticians could also play an important role in advising what can and cannot be sensibly measured. If the SDG GIF is to be useful, it is essential that users understand the limitations of these types of performance indicators. This will be especially important for donors who could conceivably make funding decisions conditional on these indicators.

The SDG indicators have hijacked, to some extent, the discussion on which statistics and data are required to support sustainable development. What has often been lost in the debate thus far is that the SDG indicators are only performance metrics – they will tell us whether a target is being achieved or not. This focus on indicators risks relegating statistics to the downstream role of monitoring and evaluation. A key role of statistics should be to inform policy decisions – this upstream or diagnostic role seems to have been, to some extent, lost in discussion. There has been relatively little debate on what additional data are required to inform and design integrated policies in order to implement actions to achieve the SDG targets. It is important that the data and statistics required to undertake risk assessments, formulate policy or design early warning systems are not forgotten during the discussions on resource mobilisation or capacity development.

9. Conclusions and Recommendations

The 2030 Agenda represents the first ever democratically forged agreement on development and will guide global development for the next ten to fifteen years. That agenda will also guide many new statistical developments and will be the driving force behind the breaking of new statistical ground.

The UN Statistical Commission and the IAEG-SDG has made tremendous progress. Despite many constraints, the SDG GIF was assembled in record time. Nevertheless, critics of the SDG indicators have criticised them for being reductionist and of watering

down the ambition of the goals and targets (Fukuda-Parr and McNeill 2019; Engle Merry 2019; Razavi 2019). Yet statistics are by definition reductionist. The question is whether in the unavoidable distillation process, the essence of the target has been faithfully captured or not. There is no question that some indicators have missed their targets and others are probably watered down from the ambitions of the target. This was almost unavoidable, given the complexity of most targets and the requirement to have only one indicator per target.

While many of the criticisms hold some water, the SDG indicator process has arguably focused more attention on global official statistics than any other UN programme. Many politicians and diplomats are now beginning to understand some of the challenges associated with consistently measuring development issues. While there are issues surrounding some indicators, they hopefully offer, at least some common ground to progress policy discussions. The SDG process also offers an opportunity for statisticians to engage and reflect with data users on the future direction of official statistics. In this context, four key strategic issues are outlined below. Without question, others could be added, but the issues highlighted here must be central to any serious discussion regarding the strategic role of NSSs and the international statistical system in the future.

1. NSOs could consider broadening their mandate to include the homologation of statistics created by third parties. Such a move would probably be welcomed by non-government organisations, civil society and academia – perhaps even the private sector, and would certainly be in keeping with the inclusive spirit of the 2030 Agenda. It would also help to maintain quality control and promote sound methodologies, transparency and openness of data (Cervera et al. 2014; Landefeld 2014; Kitchin 2015; MacFeely 2016; Hammer et al. 2017). The challenge, of course, will be for NSOs to acquire the expertise to conduct thorough and professional homologation as the frontiers of official statistics are broadening so quickly. At the global level, the United Nations could be more proactive and introduce an accreditation system (with uniform standards) that would allow unofficial compilers of statistical indicators to be accredited as 'official' for the purposes of populating the SDG GIF. One could envisage, for example, the IAEG-SDG or a similar body with the authority and competence to certify statistics as 'fit for purpose' reviewing unofficial statistics to see whether they can be certified as 'official' for the purposes of populating the SDG global monitoring framework. Without going into detail, this approach would be suitable for Tier 3 or Tier 2 indicators that otherwise run the risk of remaining unpopulated. By encouraging more active participation in the measurement, such an approach might help to domesticate the 2030 Agenda and reduce the costs of populating the SDG GIF. A detailed discussion of this idea can be found in MacFeely and Nastav (2019).
2. Newly emerging globalised digital data also offer exciting possibilities and opportunities to reconsider the national production models currently employed by NSOs and NSSs. Switching from a national to a collaborative international production model might make sense from an efficiency or international comparability perspective, but it would be a dramatic change in approach, and

possibly a bridge too far for many NSOs and governments. Globalised data are already presenting challenges as they defy national sovereignty, putting the owners and the data themselves beyond the reach of national legal systems. Governments, already struggling to enforce national laws and protect citizens, now recognise this as an important policy issue (Casalini and López González 2019; UNCTAD 2019). Nevertheless, global digital data offer opportunities to consider centralising some statistical production in a single centre, offering real international comparability, rather than replicating production many times over in individual countries. Obviously, this would not work for all domains, as issues like scale matter, as does integration with other local or national data sets. As noted above, the challenge is how to efficiently integrate the variety of data sources now available in a way that allows statistics to meet both local and global policy needs.

3. The 2030 Agenda has provided yet further justification, if further justification were needed, that countries should develop their NSSs and, to develop their NSSs and put in place a national data infrastructure (UNCTAD 2016; UNESCAP 2019). Possibly because most official statistics and disseminated administrative data are viewed as a public good, their value is not well understood or fully appreciated. Politicians do not always understand the concept of soft or nonphysical infrastructure and so may find this argument nebulous. The United Nations should take this opportunity to explain to countries that in an information age, data are an economic resource and a strategic asset, and that administrative data are not an unfortunate cost but rather a valuable national asset. Governments should also be helped to understand that data infrastructure is every bit as important as broadband or pipelines. Furthermore, the UN should emphasise that a national data infrastructure will not happen by itself, but with careful architectural design, can contribute to public sector efficiency, as well as better statistics to support public policy design and evaluation (MacFeely and Dunne 2014).
4. At the time of writing, we are almost one-third of way through the 2030 Agenda. Preparations for the post-2030 Agenda will mostly likely begin in 2028 – only seven years from now. The statistical community should prepare well in advance of the post-2030 debates. It is important that statisticians reflect on the SDG and IAEG-SDG processes and learn lessons. What worked and what didn't? How do we avoid making the same mistakes again? As a statistical community, what would we like to see changed in the follow-up programme?

This article has outlined some of the measurement challenges, tensions, unexpected consequences and strategic issues for statistics emerging from the 2030 Agenda. Again, it should be stressed that unanticipated consequences are not necessarily a bad thing. The 2030 Agenda may have inadvertently opened up new and unexpected opportunities to reimagine the traditional role of official statistics – to engage in new partnerships and build wider data ecosystems, and to develop new statistical concepts and methodologies. The ambition of the 2030 Agenda arguably provides an open door to consider bolder solutions. On the other hand, some emerging, and perhaps unexpected, clouds can also be seen on the horizon. It is not clear whether the statistical community has yet given sufficient thought to these.

10. References

Annan, K. 2000. *We the peoples: The role of the United Nations in the 21st century*. Available at: http://www.un.org/en/events/pastevents/pdfs/We_The_Peoples.pdf (accessed February 2017).

Casalini, F. and J. López González. 2019. *Trade and Cross-Border Data Flows*. OECD Trade Policy Papers, No. 220, OECD Publishing, Paris. DOI: http://dx.doi.org/10.1787/b2023a47-en.

Cervera, J.L., P. Votta, D. Fazio, M. Scannapieco, R. Brennenraedts, and T. van der Vorst. 2014. *Big Data in Official Statistics*. Eurostat ESS Big Data Event, Rome 2014 – Technical Event Report. Available at: https://ec.europa.eu/eurostat/cros/system/files/Big%20Data%20Event%202014%20-%20Technical%20Final%20Report%20-finalV01_0.pdf (accessed January 2018).

Chui, M.J., D. Farrell, S. van Kuiken, P. Groves, and E.A. Doshi. 2013. *Open Data: Unlocking innovation and performance with liquid information*. McKinsey Digital, McKinsey Global Institute. Available at: https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/open-data-unlocking-innovation-and-performance-with-liquid-information (accessed November 2019).

Cohen, B.R. 2013. "The Confidence Economy: An Interview with T. J. Jackson Lear." Available at: https://www.publicbooks.org/the-confidence-economy-an-interview-with-t-j-jackson-lears/ (accessed June 2019).

Costanza, R., L. Daly, L. Fioramonti, E. Giovannini, I. Kubiszewski, L.F. Mortensen, K.E. Pickett, V.K. Ragnarsdottir, R. de Vogli, and R. Wilkinson. 2016. "Modelling and measuring sustainable wellbeing in connection with the UN Sustainable Development Goals." *Ecological Economics* 130: 350–355. DOI: https://doi.org/10.1016/j.ecolecon.2016.07.009.

Dang, H.H. and U. Serajuddin. 2019. *Tracking the Sustainable Development Goals: Emerging Measurement Challenges and Further Reflections*. World Bank Policy Research Working Paper, No. 8843. 7 May, 2019. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3383814 (accessed May 2019).

Deen, T. 2015. "UN targets trillions of dollars to implement sustainable development agenda." Global Policy Forum. Available at: https://www.globalpolicy.org/component/content/article/271-general/52800-un-targets-trillions-of-dollars-to-implement-sustainable-development-agenda.html (accessed December 2019).

Engle Merry, S. 2019. "The Sustainable Development Goals Confront the Infrastructure of Measurement." *Global Policy* 10(1): 145–146. DOI: https://doi.org/10.1111/1758-5899.12606.

Fritz, S., L. See, T. Carlson, M. Haklay, J.L. Oliver, D. Fraisl, R. Mondardini, M. Brocklehurst, L.A. Shanley, S. Schade, U. When, T. Abrate, J. Anstee, S. Arnold, M. Billot, J. Campbell, J. Espey, M. Gold, G. Hager, S. He, L. Hepburn, A. Hsu, D. Long, J. Masó, I. McCallum, M. Muniafu, I. Moorthy, M. Obersteiner, A.J. Parker, M. Weissplug, and S. West. 2019. "Citizen science and the United Nations Sustainable Development Goals." *Nature Sustainability* 2: 922–930. DOI: https://doi.org/10.1038/s41893-019-0390-3.

Fukuda-Parr, S. and D. McNeill. 2019. "Knowledge and Politics in Setting and Measuring the SDGs: Introduction to Special Issue." *Global Policy* 10(1): 5–15. DOI: https://doi.org/10.1111/1758-5899.12604.

Hammer, C.L., D.C. Kostroch, G. Quiros, and STA Internal Group. 2017. *Big Data: Potential, Challenges, and Statistical Implications*. IMF Staff Discussion Note, SDN/17/06, September 2017. Available at: http://www.imf.org/en/Publications/SPROLLs/Staff-Discussion-Notes (accessed January 2018).

Harari, Y.N. 2018. *21 Lessons for the 21st Century*. London: Jonathan Cape.

Hulme, D. 2009. *The Millennium Development Goals (MDGs): A short history of the world's biggest promise*. BWPI Working Paper 100, Brooks World Poverty Institute, University of Manchester. Available at: https://www.unidev.info/Portals/0/pdf/bwpi-wp-10009.pdf (accessed June 2018).

IAEG-SDG (Inter-Agency and Expert Group on Sustainable Development Goal Indicators). 2018a. *Tier Classification for Global SDG Indicators*. Available at: https://unstats.un.org/sdgs/iaeg-sdgs/ (accessed June 2018).

IAEG-SDG (Inter-Agency and Expert Group on Sustainable Development Goal Indicators). 2018b. *Guidelines on Data Flows and Global Data Reporting for Sustainable Development Goals*. Background Document to Item 3 (a) of the Forty-ninth session of the United Nations Statistical Commission. Available at: https://unstats.un.org/unsd/statcom/49th-session/documents/BG-Item-3a-IAEG-SDGs-DataFlowsGuidelines-E.pdf (accessed December 2018).

IAEG-SDG (Inter-Agency and Expert Group on Sustainable Development Goal Indicators). 2019. *Data Disaggregation and SDG Indicators: Policy Priorities and Current and Future Disaggregation Plans* – Prepared by the Inter-Agency and Expert Group on Sustainable Development Goal Indicators (IAEG-SDGs). Statistical Commission, Fiftieth session. Available at: https://unstats.un.org/unsd/statcom/50th-session/documents/BG-Item3a-Data-Disaggregation-E.pdf (accessed July 2019).

IAEG-SDG (Inter-Agency and Expert Group on Sustainable Development Goal Indicators). 2019b. *Tier Classification Review – September 2019*. Available at: https://unstats.un.org/sdgs/iaeg-sdgs/tier-classification/ (accessed November 2019).

Intergovernmental Committee of Experts on Sustainable Development Financing. 2014. *Report of the Intergovernmental Committee of Experts on Sustainable Development Financing – Final Draft, 8 August 2014*. Available at: https://sustainabledevelopment.un.org/content/documents/4588FINAL%20REPORT%20ICESDF.pdf (accessed February 2017).

Jacob, A. 2017. "Mind the Gap: Analyzing the Impact of Data Gap in Millennium Development Goals' (MDGs) Indicators on the Progress toward MDGs." *World Development* 93: 260–278. DOI: https://doi.org/10.1016/j.worlddev.2016.12.016.

Jütting, J. 2016. *Capacity building, yes – but how to do it?* Available at: http://undataforum.org/WorldDataForum/capacity-building-yes-but-howto-do-it/ (accessed November 2017).

Kapto, S. 2019. "Layers of Politics and Power Struggles in the SDG Indicators Process." *Global Policy* 10(1): 134–136. DOI: https://doi.org/10.1111/1758-5899.12630.

Kitchin, R. 2015. "The opportunities, challenges and risks of big data for official statistics." *Statistical Journal of the International Association of Official Statistics* 31(3): 471–481. DOI: https://doi.org/10.3233/SJI-150906.

Landefeld, S. 2014. "Uses of Big Data for Official Statistics: Privacy, Incentives, Statistical Challenges, and Other Issues." Discussion paper presented at the United Nations Global Working Group on Big Data for Official Statistics, Beijing, China, 31 October 2014. Available at: https://unstats.un.org/unsd/trade/events/2014/beijing/Steve%20Landefeld%20-%20Uses%20of%20Big%20Data%20for%20official%20statistics.pdf (accessed January 2018).

Lebada, A.M. 2016. *Member states, statisticians address SDG monitoring requirements*. Available at: http://sdg.iisd.org/news/member-states-statisticians-address-sdg-monitoring-requirements/ (accessed December 2019).

MacFeely, S. 2016. "The continuing evolution of official statistics: Some challenges and opportunities." *Journal of Official Statistics* 32(4): 789–810. DOI: https://doi.org/10.1515/jos-2016-0041.

MacFeely, S. 2019. "The Big (data) Bang: opportunities and challenges for compiling SDG indicators." *Global Policy* 10(1): 121–133. DOI: https://doi.org/10.1111/1758-5899.12595.

MacFeely, S. and N. Barnat. 2017. "Statistical capacity building for sustainable development: Developing the fundamental pillars necessary for modern national statistical systems." *Journal of the International Association of Official Statistics* 33(4): 895–909. DOI: https://doi.org/10.3233/SJI-160331.

MacFeely, S. and J. Dunne. 2014. "Joining up public service information: The rationale for a national data infrastructure." *Administration* 61(4): 93–107. Available at: https://cora.ucc.ie/handle/10468/9513?show=full.

MacFeely, S. and B. Nastav. 2019. "You say you want a [data] revolution. A proposal to use unofficial statistics for the SDG Global Indicator Framework." *Journal of the International Association of Official Statistics* 35(3) (forthcoming).

Mair, S., A. Jones, J. Ward, I. Christie, A. Druckman, and F. Lyon. 2018. *A Critical Review of the Role of Indicators in Implementing the Sustainable Development Goals*. In: Leal Filho, W, (ed.) *Handbook of Sustainability Science and Research*. Manchester, UK. Springer, 41–56.

Muller, J.Z. 2018. *The Tyranny of Metrics*. USA: Princeton University Press, Princeton.

PARIS21. 2015. *A road map for a country-led data revolution*. Available at: http://www.oecd-ilibrary.org/docserver/download/4315051e.pdf?expires=1457406953&id=id&accname=guest&checksum=6B4747834B1E459F5E186E65EE1034B5 (accessed January 2017).

PARIS21. 2017. *Partner Report on Support to Statistics – Press 2017*. Available at: https://paris21.org/sites/default/files/2017-10/PRESS2017_web2.pdf (accessed November 2017).

Razavi, S. 2019. "Indicators as Substitute for Policy Contestation and Accountability? Some Reflections on the 2030 Agenda from the Perspective of Gender Equality and Women's Rights." *Global Policy* 10(1): 149–152. DOI: https://doi.org/10.1111/1758-5899.12633.

Runde, D. 2017. *The Data Revolution in Developing Countries Has a Long Way to Go Forbes*. 25 February, 2017. Available at: https://www.forbes.com/sites/danielrunde/2017/02/25/the-data-revolution-in-developing-countries-has-a-long-way-to-go/#620717201bfc (accessed September 2018).

Slotin, J. 2018. *What Do We Know About the Value of Data?* Global Partnership for Sustainable Development Data. Available at: http://www.data4sdgs.org/sites/default/files/services_files/Value%20of%20Data%20Report_Final_compressed_0.pdf (accessed November 2019).

The Economist. 2015a. "The 169 commandments." 28 March 2015. *The Economist*. Available at: https://www.economist.com/leaders/2015/03/26/the-169-commandments (accessed December 2019).

The Economist. 2015b. "Assessing development goals: The good, the bad and the hideous." 26 March 2015. *The Economist*. Available at: https://www.economist.com/international/2015/03/26/the-good-the-bad-and-the-hideous (accessed December 2019).

UN (United Nations). 2013. *Open Working Group proposal for Sustainable Development Goals*. Available at: https://sustainabledevelopment.un.org/content/documents/1579SDGs%20Proposal.pdf (accessed February 2017).

UN (United Nations). 2015a. *Transforming our world: the 2030 Agenda for Sustainable Development Resolution 70/1* adopted by the General Assembly on 25 September 2015. Available at: http://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=E (accessed February 2017).

UN (United Nations). 2015b. *Agenda, 2030 "to-do list for people and planet"*, Secretary-General Tells World Leaders Ahead of Adoption [press release]. Available at: http://www.un.org/press/en/2015/sgsm17111.doc.htm (accessed January 2015).

UN (United Nations). 2000. *United Nations Millennium Declaration*. Resolution 55/2 adopted by the General Assembly. Available at: http://www.un.org/en/ga/search/view_doc.asp?symbol=A/RES/55/2 (accessed February 2017).

UNCTAD (United Nations Conference on Trade and Development). 2016. *Development and globalisation: Facts and Figures 2016*. Available at: http://stats.unctad.org/Dgff2016/index.html (accessed October 2016).

UNCTAD (United Nations Conference on Trade and Development). 2019. *Digital Economy Report 2019 – Value Creation and Capture: Implications for Developing Countries*. Available at: https://unctad.org/en/pages/PublicationWebflyer.aspx?publicationid=2466 (accessed November 2019).

UN (United Nations) Department of Economic and Social Affairs. 2019. Graduation from the LDC category: Timeline – Country Graduations. Available at: https://www.un.org/development/desa/dpad/least-developed-country-category/ldc-graduation.html (accessed December 2019).

UNECE (United Nations Economic Commission for Europe). 2016. *Outcomes of the UNECE Project on Using Big Data for Official Statistics*. Available at: https://statswiki.unece.org/display/bigdata/Big+Data+in+Official+Statistics (accessed February 2018).

UNECE (United Nations Economic Commission for Europe). 2019. *Recommendations for Promoting, Measuring and Communicating the Value of Official Statistics*. Available at: http://www.unece.org/index.php?id=51139 (accessed November 2019).

UN (United Nations) Environmental Programme. 2015. *Clarifying terms in the SDGs: Representing the meaning behind the terminology*. Available at: http://unstats.un.org/sdgs/files/meetings/iaeg-sdgs-meeting-02/Statements/UNEP%20-%20Clarifying%20terms%20in%20the%20SDGs.pdf (accessed January 2017).

UNESCAP (United Nations Economic and Social Commission for Asia and the Pacific. 2019. *Integrated Statistics: A journey worthwhile*. Stats Brief. July 2019, Issue No. 19. Available at: https://www.unescap.org/sites/default/files/Stats_Brief_Issue19_Jul2019_Integrated_Statistics.pdf (accessed November 2019).

UN (United Nations) General Assembly. 2017. *Resolution adopted by the General Assembly on 6 July 2017 – Work of the Statistical Commission pertaining to the 2030 Agenda for Sustainable Development*. Seventy-first session. A/RES/71/313. Available at: https://undocs.org/A/RES/71/313 (accessed June 2019).

UN (United Nations) Statistical Commission. 2017. *Report on the forty-eighth session (7–10 March 2017)*. Economic and Social Council, Official Records 2017 – Supplement No. 4, E/2017/24-E/CN.3/2017/35. Available at: https://undocs.org/pdf?symbol=en/E/2017/24.

Warren, S. 2015. Analyzing the #SDGs on Twitter: The social media response to the Sustainable Development Goals. Available at: https://www.owler.com/reports/dalberg-global/dalberg-global-blog-analyzing-the–sdgs-on-twitter/1449841459497 (accessed December 2019).

Received June 2019
Revised November 2019
Accepted December 2019

sciendo

Journal of Official Statistics, Vol. 36, No. 2, 2020, pp. 379–409, http://dx.doi.org/10.2478/JOS-2020-0020

Explaining Inconsistencies in the Education Distributions of Ten Cross-National Surveys – the Role of Methodological Survey Characteristics

Verena Ortmanns[1]

Surveys measuring the same concept using the same measure on the same population at the same point in time should result in highly similar results. If this is not the case, this is a strong sign of lacking reliability, resulting in non-comparable data across surveys. Looking at the education variable, previous research has identified inconsistencies in the distributions of harmonised education variables, using the International Standard Classification of Education (ISCED), across surveys within the same countries and years. These inconsistencies are commonly explained by differences in the measurement, especially in the response categories of the education question, and in the harmonisation when classifying country-specific education categories into ISCED. However, other methodological characteristics of surveys, which we regard as 'containers' for several characteristics, may also contribute to this finding. We compare the education distributions of nine cross-national surveys with the European Union Labour Force Survey (EU-LFS), which is used as benchmark. This study analyses 15 survey characteristics to better explain the inconsistencies. The results confirm a predominant effect of the measurement instrument and harmonisation. Different sampling designs also explain inconsistencies, but to a lesser degree. Finally, we discuss the results and limitations of the study and provide ideas for improving data comparability.

Key words: Comparative research; cross-national surveys; survey characteristics; education.

1. Introduction

Education is a key socio-demographic variable that is measured in nearly every survey (Smith 1995). Education is central in social stratification research, for instance, when analysing educational inequalities and how social class origin affects education (Breen and Jonsson 2000, 2005; Müller and Karle 1993), or when analysing returns to education, for example how education determines individuals' income and socio-economic status

[1] GESIS – Leibniz Institute for the Social Sciences, Department Survey Design and Methodology, P.O. Box 12 21 55, 68072 Mannheim, Germany. Email: verena.ortmanns@gesis.org

Acknowledgments: I especially thank Silke Schneider, Michael Braun, Matthias Sand, Roberto Briceno-Rosas and Patricia Hadler for their helpful inputs and suggestions on this study as well as their feedback on earlier versions of this article. I also would like to thank the editor and especially the four anonymous reviewers for their remarks which helped to improve this article. I also thank my student assistants, Tim Dennenmoser, Svenja Friedel, Zeynep Umuc Demirci, Paweł Komendziński and Kimberly Herbst for their help with researching, collecting and coding the survey characteristics. I would also like to thank the UK Data Archive, Eurofound, the Norwegian Centre for Research Data, the GESIS Data Archive and the German Microdata Lab at GESIS for providing the data sets, the documentation of the survey characteristics and for their support. Finally, I thank Jane Roberts for proofreading. This work was supported by the Leibniz Association through the Leibniz Competition (formerly SAW Procedure) grant number SAW-2013-GESIS-5, as well as the European Union's Horizon 2020 research and innovation programme under grant agreement No 654221.

(Becker 1993; Blau and Duncan 1967; Bol and Van de Werfhorst 2013). Outside of stratification research, the education variable is an important proxy variable for another concept, such as cognitive competencies, and it is a widely used as a background or control variable. Quite often studies find a substantial effect of the education variable, for example when analysing values and behaviours, such as, political attitudes or voting behaviours (Bekhuis et al. 2014; Weakliem 2002), gender role attitudes (Bolzendahl and Myers 2004; Kalmijn 2003) or attitudes towards minorities and immigrants (Coenders and Scheepers 2003; Semyonov et al. 2008; Hyman and Wright 1979). In survey methodological research, the education variable is important because together with sex and age, it is often used to assess the comparability of the survey data, for instance with official data sources (Peytcheva and Groves 2009). Furthermore, education is often included when calculating post-stratification weights, which aim to correct for non-sampling errors such as nonresponse and may decrease the variance of a survey's estimate (e.g., ESS 2014b). Clearly the education variable is important for different purposes, and ideally should be of high measurement quality.

Previous research compared the education distribution across surveys within countries and years to assess how reliable the distribution of education is measured across surveys and thus how comparable the data are. For identical populations and time points, one would expect only minimal variation in the data. However, studies repeatedly revealed inconsistencies in education distributions across surveys even when they use the same harmonised education variables (Kieffer 2010; Ortmanns and Schneider 2016a, 2016b; Schneider 2009). These discrepancies indicate that the data cannot be comparable in some way. However, especially for cross-national comparative research, data need to be comparable. In more detail, the study of Kieffer (2010) observed discrepancies when comparing the distribution for the European Social Survey (ESS) with the EU-LFS for France. Large deviations were identified for the first three waves of the ESS in 2002, 2004 and 2006; while for 2008, the deviation was smaller. Schneider (2009), who compared data from 2002 to 2007, also identified inconsistencies when comparing the distributions for most countries in the European Union Statistics on Income and Living Conditions (EU-SILC), and in the ESS with the EU-LFS. Ortmanns and Schneider (2016b) replicated and extended this work by comparing education distributions for European countries included in four public opinion surveys between 2008 and 2012. They analysed the Eurobarometer, the European Values Study EVS, the International Social Survey Programme (ISSP) and the ESS, which was used as the reference survey. In the most comprehensive study to date, Ortmanns and Schneider (2016a) analysed seven cross-national survey programmes, again looking at the period 2008 to 2012. They included OECD's Programme for the International Assessment of Adult Competencies (PIAAC), EU-SILC, Eurobarometer, ESS, EVS and ISSP, and compared the education distributions for the same countries and years to the respective distribution in the EU-LFS. Since this study is the basis for this article, we will briefly summarise the main result to illustrate the problem. Ortmanns and Schneider (2016a) found that on average, 13% of respondents would have to change education categories to achieve an equal distribution with the EU-LFS. They also found substantial variation across surveys, ranging from 1% to almost 50%. These inconsistencies cannot reflect actual differences in the education distribution because it should be rather stable for the same country and

year. Instead, these inconsistencies indicate a severe problem with data comparability across surveys, and thus methodological differences between the surveys must explain the observed deviations.

To date, researchers explain those inconsistencies commonly by differences in the measurement of education or the way country-specific response categories are classified into the International Standard Classification of Education (ISCED) (Kieffer 2010; Ortmanns and Schneider 2016a, 2016b). However, we cannot be sure that these are the only or most important factors just because they can be observed easily and are reported more often. Ortmanns and Schneider (2016a) identify single cases where they hypothesise that differences in the survey characteristics such as data collection modes, sampling designs, as well as selective unit nonresponse might also explain the inconsistencies because they do not find any problem in the measurement or the assignment of ISCED codes. Those survey characteristics refer to methodological aspects of a survey, and they differ across surveys because they are designed and organised differently, and apply different methodological standards. Thus, the survey characteristics influence the quality of the survey and its data. To systematically analyse and test the impact of surveys' methodological characteristics, we need an in-depth, quantitative and comprehensive analysis.

Such an analysis is conducted in this study, which analyses the impact of 15 survey characteristics and how they contribute to inconsistent education distributions across surveys within countries and years. As a starting point, we use the results from Ortmanns and Schneider (2016a), comparing the education distributions of six surveys with the EU-LFS for the years 2008 to 2012. We further extend the range of cross-national surveys by adding the Adult Education Survey (AES), the European Quality of Life Survey (EQLS), and the European Working Condition Survey (EWCS). Hence, this study compares the education distributions often cross-national surveys for 31 European countries. The research question is: Can survey characteristics explain the inconsistencies identified in the education distributions across surveys? Thirteen hypotheses are formulated and tested by estimating regression models.

Section 2 describes these cross-national surveys and how they measure education. It also introduces the challenges of comparing the education distributions and the survey characteristics across surveys. In Section 3, we present several different survey characteristics and derive our hypotheses regarding their contribution to the inconsistencies in education distributions. We use the Total Survey Error (TSE) framework (Groves et al. 2009; Groves and Lyberg 2010) to structure this presentation. In Section 4, the variables and methods are described, before presenting the results in Section 5. In Section 6, we discuss the results and limitations of the study and provide ideas for improving data comparability.

2. The Cross-National Surveys and their Education Measures

2.1. The Cross-National Surveys Covered in this Study

This study compares the education distributions of nine large-scale, cross-national surveys to the EU-LFS (Eurostat 2008, 2010a, 2011b, 2012), which we use as a benchmark, and estimates the impact of survey characteristics on the observed inconsistencies in the

education distributions. To better understand the challenges of estimating the impact of survey characteristics when using the EU-LFS as a benchmark, and the consequences for the design of this study, we start with a brief description of the survey programmes.

Since the beginning of the EU-LFS in the 1970s, it has provided official household data for monitoring employment and unemployment in all EU countries and some European non-EU countries. The large number of countries included in the survey, the large sample sizes, the relatively high response rates and the probability-based sampling should produce representative high-quality data and thus an accurate education distribution for each country. Furthermore, the EU-LFS provides annual data, is fairly well documented, and it applies the official ISCED mappings. Thus, it is the most authoritative source regarding education data in Europe. Statistics based on the EU-LFS are, for instance, used in the annual OECD reports "Education at a Glance" (e.g., OECD 2015, 2016, 2017). EU-LFS data are also used when defining goals of the Europe 2020 strategy to enhance participation in education in all European countries (Eurostat 2019). The distribution of the EU-LFS education variable is also used as reference for other surveys, such as the ESS, when comparing or weighting data (ESS 2014a, 2014b). We are not aware of another official cross-national survey that fulfils all these criteria. Census data, for instance, typically do not provide harmonised data, which can be used for international comparisons; those have to be generated by the researcher herself. More important, to our knowledge, researchers cannot simply access an integrated data set of the latest official census data for all European countries. Hence, we use the EU-LFS as the benchmark survey in this study.

However, the EU-LFS also does not reflect the 'true' education distributions of the countries. The EU-LFS is an output harmonised survey, meaning the national surveys, to a large extent, are independent of each other and follow different national regulations. This applies for nearly all survey characteristics. Survey participation, for instance, is mandatory in roughly half of the countries the EU-LFS, but it is voluntary for the other countries. The response rate also varies greatly across countries between 30% and 98%. Furthermore, the countries use different sampling designs (simple or complex designs), as well as different modes of data collection (face-to-face, telephone, self-administered or mixed-mode). Of course, some guidelines and rules are specified to achieve as much comparable statistics as possible across countries, but the national survey designs entail quite different survey characteristics across the countries participating in the EU-LFS. This considerable variation in the survey characteristics of the EU-LFS forces us to analyse the impact of these survey characteristics with a rather broad approach. Therefore, we cannot assess which data collection mode causes more or fewer inconsistencies in the education distribution. Instead, we can only analyse whether mode differences between the survey in question and the EU-LFS affect the education distribution. As indicated, this applies to all survey characteristics; thus, we can only assess whether differences in the survey characteristics can contribute to inconsistencies in the education distributions across surveys within the same countries and years. This has to be considered when developing the hypotheses, and it adds complexity when operationalising the variables and interpreting the results. Nevertheless, it is important to mention that for all surveys, good documentation of the survey characteristics is an essential precondition for this study to identify how the survey characteristics differ across surveys within the same countries and years.

Another official survey included in this analysis is the EU-SILC (Eurostat 2010b). It was launched in 2003 with the aim of providing cross-sectional and longitudinal official micro-data on income, poverty, social exclusion, as well as living and housing conditions in the EU. We also analyse data from PIAAC (OECD 2013) and the AES (Eurostat 2011a), which focus on education. PIAAC is an OECD survey that measures adults' general basic skills, and first collected data in 2011/12 across OECD countries. The AES is a Eurostat survey that covers participation in formal and non-formal education and training of adults in EU countries. It began in 2007 and has been repeated nearly every fifth year. We also analyse data of the Eurobarometer (European Commission 2012), which was set up by the European Commission in the 1970s to monitor public attitudes towards the EU and related topics in all Member States. So far, the ISCED classification has only been implemented in three Eurobarometer studies, two of them have been conducted in 2010 and one in 2011. Additionally, we also analyse date from the EQLS (Eurofound 2014) and the EWCS (Eurofound 2011). Both surveys include all EU countries and they are funded through Eurostat and realised by Eurofound. The EQLS is conducted every four to five years since it was launched in 2003. The survey questions European citizens on general circumstances of their lives, such as employment, income, housing, family, happiness, and well-being. The EWCS was launched in 2005 and also runs quinquennially. It focuses on different aspects of employment, such as working time, learning and training, earnings and financial security, as well as work-life balance and health.

Lastly, three data sources from the academic community are included that cover different topics related to individuals' attitudes, beliefs, values and behaviour: the ESS (ESS 2016a, 2016b, 2016c), the EVS (EVS 2016), and the ISSP (ISSP Research Group 2015, 2016). The ESS was set up in 2002 and runs every second year in around 30 European countries. The EVS was launched in 1981, and data from five rounds of the survey are now available. The ISSP is an annual survey set up in 1985, and like PIAAC, it extends beyond Europe.

These surveys partly differ in the definition of their target population, for instance with regard to age groups. To render the samples as comparable as possible, we include only respondents aged 25 to 64 in all surveys. The EWCS focuses on people who are employed and thus, we restrict the analytic sample of the EU-LFS to employed respondents when comparing it to the EWCS.

2.2. Measuring and Comparing Educational Attainment in Cross-National Surveys

Asking respondents about their educational attainment is standard in almost all surveys in the social sciences. This question often refers to individuals' highest formal qualification or their highest completed educational level for which a diploma or certificate from a school, a formal vocational training or an institution of higher education or university is awarded. Respondents usually answer this question by selecting a category from a list. Those lists are necessarily country-specific, as education systems differ in their institutions and the names of the qualifications, which cannot be accurately translated (Braun and Mohler 2003; Schneider et al. 2016). Therefore, the ex-ante output harmonisation approach (Ehling 2003) is commonly used in cross-national surveys. Before data collection, the survey teams agree on a standard classification or a coding scheme and

ideally set up guidelines specifying what has to be considered when developing the country-specific answer categories. The mapping of these categories to the standard classification, which is used to compare education across countries, is also developed in advance (Ehling 2003; OECD and Eurostat 2014). To harmonise the education categories across countries, most surveys choose the ISCED classification. This was designed by UNESCO in the 1970s and revised in 1997 and 2011. It aims to enable comparisons of country-specific education programmes for producing international education statistics. The ISCED classification defines international levels and types of education (UNESCO-UIS 2006), and education ministries and national statistical institutes map their educational programmes and qualifications to it. The most recent version of the classification was not yet implemented in most surveys for the years analysed, thus limiting this research to ISCED 97.

The main levels of ISCED 97 are:

- ISCED 0: Pre-primary education (or not completed primary education)
- ISCED 1: Primary education or first stage of basic education
- ISCED 2: Lower secondary or second stage of basic education
- ISCED 3: Upper secondary education
- ISCED 4: Post-secondary non-tertiary education
- ISCED 5: First stage of tertiary education
- ISCED 6: Second stage of tertiary education.

The focus here is on comparing the main levels of ISCED 97, ignoring the additional complementary dimensions on programme orientation, destination, duration and position in the national qualification structure, as most of the surveys analysed do not use them. All surveys we analysed implement the main levels of the ISCED classification or a variant thereof, from which we can derive the main level of ISCED 1997 for comparing the distributions. We need to aggregate ISCED levels 0 and 1 and levels 5 and 6 because those categories are not separated in all surveys. When comparing the EU-LFS and the ISSP, we also need to aggregate ISCED levels 3 and 4 (see Tables S1 and S2 in the online Supplemental material).

Following Ortmanns and Schneider (2016b, 2016a), we calculate Duncan's Dissimilarity Index (Duncan and Duncan 1955) to compare the education distributions between the EU-LFS, used as the benchmark survey, and the other surveys, which also use the ISCED classification. The index is defined as: $D = \frac{1}{2}\sum_{i=1}^{k}|x_i - y_i|$ where x_i denotes the number of observations in category i out of k ISCED categories for country A in year B in survey S, and y_i denotes the same for country A in year B in survey T. To interpret the resulting numbers as percentages, the index is rescaled to range from 0 to 100. This tells us how large the percentage is that needs to change categories to achieve equal education distributions between the EU-LFS and the survey in question.

Figure 1 shows the summary statistics of Duncan's Dissimilarity Index when comparing the education distributions between the EU-LFS and the other surveys within the same countries and years. The exact values can be found in Table S3 in the Supplemental material; these are used later as the dependent variable. We observe the smallest value of 1% in Duncan's index when comparing data for the Czech Republic from the 2010 EU-LFS and EU-SILC; this indicates nearly perfectly consistent data. The largest

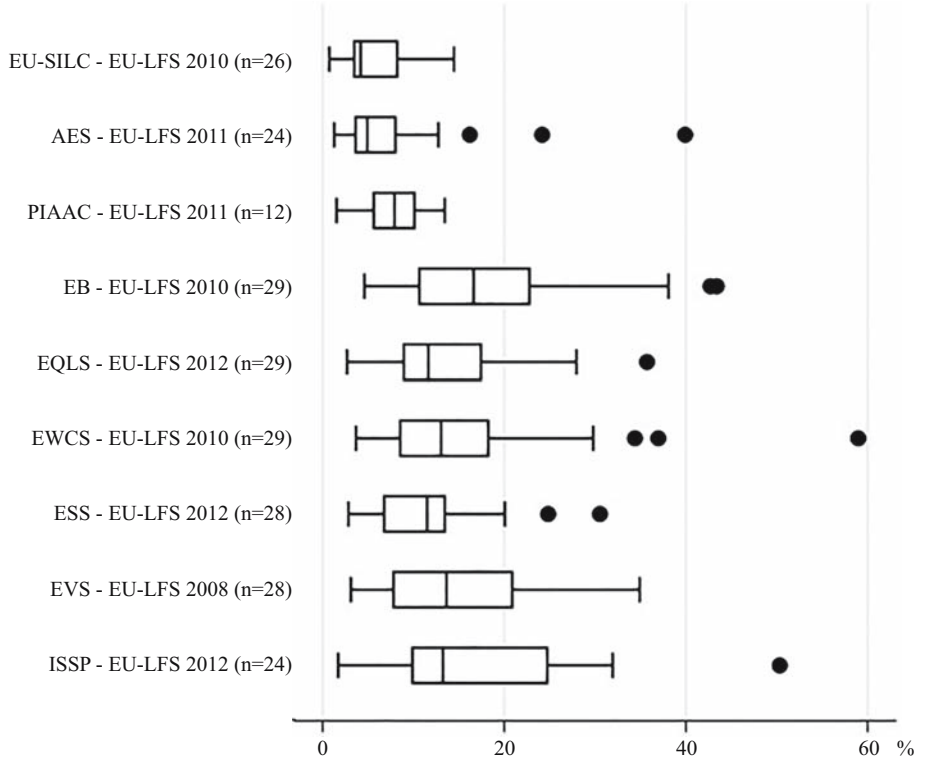


Fig. 1. Boxplots of Duncan's Dissimilarity Index across countries for all survey comparisons.
Notes: Here 'n' indicates the number of countries included in the analyses. Data sources, see in online Supplemental material.

deviation of 59% is found when comparing EU-LFS and EWCS data for Germany from 2010, which is even higher than the highest deviation identified by Ortmanns and Schneider (2016a). Overall, the mean inconsistency is almost 13%, meaning that on average 13% of respondents would need to change categories to achieve a distribution equal to that in the EU-LFS, which is the same result as found by Ortmanns and Schneider (2016a) based on a more limited set of international surveys. Duncan's Dissimilarity Index should, however, be close to zero because the education distributions should not vary across surveys when analysing the same country and year. This is clearly not the case. Looking at the individual surveys, we find the lowest discrepancy of roughly 6% when comparing the education distributions of the EU-LFS and the EU-SILC. When comparing the distributions of PIAAC and the AES to the EU-LFS, the discrepancy is 8%. We interpret these deviations as relatively small because they are clearly below the mean value of 13%. Duncan's index indicates a discrepancy of 12% between the ESS and the EU-LFS, 14% between the EQLS and the EU-LFS and 15% between the EVS and the EU-LFS. These percentages are around the mean value (between 10 and 15%) and, thus, we regard those as intermediate discrepancies. The comparison between the EWCS and the EU-LFS indicates a discrepancy of 16% and between the ISSP and the EU-LFS the discrepancy is 17%. We find the largest discrepancy of 19% when comparing the education distributions

of the EU- LFS and the Eurobarometer. We interpret these deviations, which are above 15%, as larger inconsistencies.

3. Survey Characteristics

In order to explain differences between surveys, countries and years in terms of how well their education distribution matches that produced by the EU-LFS for the respective country and year, we refer to the Total Survey Error framework (Groves et al. 2009; Groves and Lyberg 2010) that describes different sources of errors that can appear at different stages of a survey. We use this framework for structuring the survey characteristics according to the different error sources, following the dimensions of representation of the population and measurement. An overview of all survey characteristics analysed in this study can be found in Table 1.

Considering that all surveys we analysed in this study are cross-national, we have to be aware that the survey characteristics do not only vary across surveys, but also across participating countries (Kohler 2008; Słomczyński et al. 2016). Different errors in the countries also reduce quality in terms of comparability across countries and/or surveys, as described in the application of the TSE approach to cross-national surveys (Smith 2010, 2011).

Some methodological survey characteristics are design features of the survey that can be changed in principle, such as the mode of data collection or fieldwork duration. Other survey characteristics, such as response rate, cannot be changed directly by the survey organisers. In methodological studies, the relationship between different kind for survey characteristics haven been examined as well as the impact of single characteristics on the data quality. For instance, studies have assessed whether the mode of data collection or offering incentives have an impact on response rates (Church 1993; Daikeler et al. 2019). Other studies evaluate the representation of the population of cross-national surveys by systematically comparing single survey characteristics across countries for a single survey (Kaminska and Lynn 2017) or across several surveys (Kohler 2007). Based on this research, best practice guidelines for survey organisers are developed (see e.g., Groves and Couper 1998, chap. 11).

3.1. Survey Characteristics Related to the Representation of the Population

In this section, we present several survey characteristics related to the representation of the population and how they could, theoretically, explain the inconsistencies in the education distributions between the EU-LFS, our benchmark, and the survey in question. When developing our hypotheses on the impact of the survey characteristics, we have to consider that those differ across countries also for the EU-LFS (see Subsection 2.1). Thus, we will only formulate undirected hypotheses indicating that differences in the survey characteristics of the EU-LFS and the survey in question might explain discrepancies in the education distributions across surveys within the same country and year.

Looking at the dimension of representation in the TSE approach, four kinds of errors are distinguished: coverage, sampling, unit nonresponse, and adjustment error (Groves et al. 2009). Coverage error emerges at an early stage even before drawing a sample; it arises when there is a discrepancy between the sampling frame and the target population.

Table 1. Overview of the survey characteristics and their operationalisation.

Dimension and errors of the TSE		Survey characteristic	Values	Values when comparing with EU-LFS
Representation of the population	Sampling error	Sampling design	Simple, complex	0 = equal, 1 = unequal
		Final sampling unit	Individual, household, dwelling/address	0 = equal, 1 = unequal
		Sample size	n	Absolute difference in the sample size divided by 1000
	Nonresponse error	Response rates	In percent	0 = equal response rate, 1 = higher, < 30 percentage points, 2 = lower, < 30 percentage points, 3 = lower, ≥ 30 percentage points, 4 = not available
		Survey participation	Mandatory, voluntary	0 = equal, 1 = unequal
		Fieldwork duration	Days	0 = equal duration, 1 = shorter, < 90 days, 2 = longer, > 90 days, 3 = longer, ≥ 90 days
	Sampling and nonresponse error	Index to validate probability sampling	Chance of interviewing a man/woman of a married couple living together in a two-person household	0 = equal, 1 = unequal
		Index on gender and age	Distribution of men and women for following age groups: 25–34, 35–44, 45–54, 55–64	Deviations in percent, indicating differences in the gender and age distribution

Table 1. Continued.

Dimension and errors of the TSE		Survey characteristic	Values	Values when comparing with EU-LFS
Measurement	Measurement error	Response categories of the education question		0 = same, 1 = similar, 2 = different
		Proxy-reporting	Yes, no	0 = equal, 1 = unequal
		Information taken from register	Yes, no	0 = equal, 1 = unequal
	Processing/ harmonisation error	Applying official ISCED mapping	Official ISCED mapping is applied, intended deviation, accidental deviation	0 = equal, 1 = unequal
		Degree of centralisation when applying ISCED	Decentralised, partly centralised, entirely centralised	0 = equal, 1 = unequal
Representation & measurement	Sampling, nonresponse, measurement and processing error	Mode of data collection	Face-to-face, telephone, self-administered, mixed-mode	0 = equal, 1 = unequal
		Fieldwork agency	Institute of public authority, university/scientific institute, commercial institute	0 = equal, 1 = unequal

Sampling error occurs when randomly taking a subset of sampling units from the sampling frame. When assessing sampling error, it is important to notice that most surveys analysed here use probability-based sampling methods, but that in the last stage, random-route approaches are applied in a few surveys. The survey characteristics on the sampling design and the final sampling unit reflect both coverage and sampling error and sample size only sampling error.

The *sampling design* influences the composition of the sample and thus also the education distribution. Almost every sampling design excludes some people from the target population, which might cause under- or over-coverage of certain groups (Groves and Couper 1998; Lohr 2009). In this article, we only distinguish between simple and complex sampling designs. In a simple design, the respondent is selected directly from an official register by means of a simple random sample. This is usually the case in the Scandinavian countries, which have central population registers. Ten countries of the EU-LFS have such a sampling design. In contrast, a complex sampling design might also use an official register, but multiple stages are used in the selection process. Other examples of a complex design are random digit dialing, and those where in the final stage a random route technique is applied. If the sampling design differs between the EU-LFS and the other survey, differences in the sample composition are likely, which might contribute to inconsistencies in the education distributions across surveys within the same countries and years (Hypothesis 1). Differences in the sample composition can also occur when both surveys apply complex sampling designs that differ from each other, for example through using different sampling frames. Unfortunately, generating a more detailed differentiation, for example by including additional information on the sampling frame, was not possible due to unstandardised or lacking information. For instance, it was also not possible to consider the information on how the surveys deal with institutionalised population because this often is not a central aspect in the documentation, although it is important to better assess errors in coverage and sampling (Schanze 2017).

Next, we look at the *final sampling unit*. We differentiate between an individual, a household or a dwelling/address. In most countries, the EU-LFS and the EU-SILC are household surveys and the dwelling/address or the household are the final sampling unit. Usually, in those surveys all respondents in a household above a specified age (15 in the EU-LFS, 16 in the EU-SILC), and more than one respondent at the same address or dwelling, are interviewed. This increases the chance of being selected to answer the questionnaire. In contrast, most other surveys use the individual respondent as the final sampling unit, and the individual probability of being selected is lower in these surveys (Groves et al. 2009). The different selection probabilities can influence the sample compositions and thus also the education distribution. To not overestimate the effect of the different sampling units, especially for the household surveys, data are weighted using available design weights. Therefore, we hypothesise that differences in the final sampling units across surveys might not affect the inconsistencies in the education distributions across surveys (Hypothesis 2).

The *sample size* of a survey matters because previous research shows that surveys with a larger sample size are more accurate, as the sampling error decreases (Biemer and Lyberg 2003). Surveys with smaller samples are more likely to have a sampling error that can lead to a slightly different sample composition and thus to a slightly different education

distribution. All analysed surveys have rather large samples; however, the EU-LFS has by far the largest sample size for each country. Thus, we will definitely observe deviations in the sample size across the surveys. However, we estimate that these differences in the sample size might not contribute to the discrepancies in the education distribution (Hypothesis 3).

The nonresponse error, focusing on unit nonresponse, results in lacking representativeness of the sample. This error occurs if respondents systematically differ from nonrespondents, that is sample members who refuse to participate in the survey or who cannot be interviewed. Here, we look at the following survey characteristics: mandatory survey participation, fieldwork duration and response rate. The survey characteristic on *mandatory survey participation* indicates that respondents are forced to participate in the survey. Usually, those surveys achieve higher response rates, and the nonresponse error is low because respondents who would refuse in voluntary surveys are often included in mandatory ones. Thus, we hypothesise that differences in mandatory survey participation across the EU-LFS and the other surveys might explain inconsistencies in the education distribution (Hypothesis 4). In the analysed surveys, participation is mandatory for only a small number of countries and surveys, namely 13 countries in the EU- LFS and nine in the AES.

Regarding fieldwork duration previous research indicate that longer field periods increase the chance of contacting and interviewing hard-to-reach respondents, whereas shorter fieldwork durations often leave less time for follow-ups. Thus, for surveys having a shorter fieldwork duration, errors of nonresponse become more likely (Biemer and Lyberg 2003). In the EU-LFS, fieldwork duration is usually three months and we distinguish whether the fieldwork compared to the EU-LFS is longer or shorter. We expect that different fieldwork durations—either considerably shorter or considerably longer than the benchmark—might increase inconsistencies in the education distribution across surveys within the same countries and years (Hypothesis 5).

The *response rate* is an important quality indicator and survey organisers invest a great deal of money in increasing it, for instance, by offering incentives to the respondents (Singer and Ye 2013; Groves et al. 2006). The response rate of the EU-LFS is relatively high, due to mandatory survey participation in some countries and because proxy-reporting is generally permitted. In contrast, for most other surveys the response rates are much lower and this might indicate that their realised samples can differ from the sample of the EU-LFS, that is, there is a higher risk of nonresponse error. Thus, we hypothesise that large differences in the response rates between the EU-LFS and the other surveys within countries and years could contribute to explaining inconsistencies in the education distributions (Hypothesis 6). However, we know that a high response rate alone is not enough to avoid nonresponse error (Bethlehem et al. 2011; Groves and Peytcheva 2008). Nevertheless, we decided to include this survey characteristic because we have no better indicator of the nonresponse bias.

The last error related to representation of the population is adjustment error. It emerges after data collection when calculating weights. This error is not taken into account in this study, because data are only weighted using design weights that correct for different inclusion probabilities due to different sampling designs across countries. Applying post-stratification weights that also correct for nonresponse errors is not feasible because those often correct for education, frequently by using the distribution of the EU-LFS as

benchmark (e.g., ESS 2014b). This would lead to an (almost) equal distribution of the two surveys that are being compared.

Some specifications of the described survey characteristics relating to representation of the population are rather broad, for instance regarding the sampling design and sampling unit. This is caused by vague and sometimes also questionable documentation, particularly the design of the sampling process (for more information on the different standards in documentation, see Kohler 2008; Słomczyński et al. 2016). Therefore, it is advisable to also look directly into the data and check the realised representation. Firstly, we generate Sodeur's Index to validate probability sampling of the survey (Sodeur 1997, 2007). This index is based on the assumption that in a random sample, the chance of interviewing a man or a woman in a married couple living together in a two-person household is equal, namely 50:50. We adapt this and define the observed distribution of the EU-LFS as a benchmark. For calculation, we firstly restrict all samples to the 25 to 64 age group and married couples living in two-person households. Unfortunately but not unexpectedly, the required variables on marital status and household composition differ greatly across surveys, so adaptations are needed (for details see Annex 1 in Supplemental material). We calculate the gender distribution of this restricted sample and compare it to the distribution identified in the respective sample of EU-LFS, applying the following formula: $B_{UNR} = \frac{\hat{p}-p}{\sqrt{var(\hat{p})}}$ where p is the proportion of women in the EU-LFS and $\hat{p}$ is the proportion of women in the survey in question for the same country and year. Finally, the 95% confidence interval is calculated so we can decide whether the gender distribution between the EU-LFS and the other survey is equal or not within the same country and year. Secondly, we calculate an index to compare the gender and age distributions for four age groups (25–34, 35–44, 45–54, and 55–64) across surveys. Here, we again calculate Duncan's Dissimilarity Index (Duncan and Duncan 1955) and we use the distribution of the EU-LFS as benchmark.

3.2. Survey Characteristics Related to Measurement

On the measurement dimension of the TSE framework, there are three kinds of error that can occur: invalidity, measurement error, and processing error (Groves et al. 2009). Invalidity occurs when there is a disparity between the theoretical construct (what is intended be measured) and what is actually measured by the indicator. In this study, we do not expect to find invalidity because every survey asks respondents for their highest educational attainment in an equivalent way, asking respondents for their highest certificate/degree or their achieved educational level.

Measurement error occurs when a mismatch exists between the ideal measurement and the actual response obtained from the respondent. A potential source of measurement error across surveys is differences in the *response categories* in the education question. Previous research shows many examples pointing at differences in the measurement instrument as a source of inconsistent education data (Kieffer 2010; Ortmanns and Schneider 2016a, 2016b; Schneider 2009). For instance, when surveys use ambiguous terms or generic descriptions of educational qualifications, instead of the official name of the qualifications, the chance that the response categories differ across surveys is quite high. Thus, this survey characteristic seems to be of some importance when explaining inconsistencies in

the education distributions. In the education question, the response categories are the key element influencing respondents' answers. All analysed surveys use country-specific response categories for the education question. To assess the similarity of the response categories of the EU-LFS and the other surveys, we qualitatively compared the education categories for every survey, country and year and generated an index. It distinguishes whether the categories are the same as, similar to, or different from the categories used in the EU-LFS. Detailed information on this index is provided in Annex 2 (Supplemental material). In general, we know that different stimuli can affect respondents' answers (Groves et al. 2009) and this also seems to occur with the education question, even though it is a factual question. Thus, different response categories are a probable explanation for inconsistencies in the education distributions (Hypothesis 7).

Relating to the measurement, we also measure whether *proxy-reporting* is allowed or prohibited. If the survey allows proxy-reporting, a respondent's partner or (adult) child might answer the questions instead of the selected respondent, or the 'head of the household' responds for every household member. Proxy-reporting can only be used in household surveys; thus, it applies to the EU-LFS, EU-SILC and the AES. Proxy-reporting is cognitively demanding, and measurement errors are likely due to lack of knowledge leading to incorrect answers (Blair et al. 2011; Kreuter et al. 2010; Moore 1988). Thus, we expect that differences in the allowability of proxy-reporting can contribute to inconsistencies in the education distribution across surveys (Hypothesis 8).

The last survey characteristic related to measurement error distinguishes *whether respondents' educational attainment is retrieved from a register* or not. Some countries, mostly Scandinavian ones, have population registers from which socio-demographic information, including education, can be directly retrieved. Register information is regarded as high quality and trustworthy (Biemer and Lyberg 2003). Therefore, differences in this survey characteristic on retrieving information from a register may explain inconsistencies in the education distribution (Hypothesis 9). However, we also have to be aware that register information is not free of errors either, due to delayed updates, especially for younger people who are currently in education (Kleven and Ringdal 2017). Only four countries of the EU-LFS use register information.

Next, we look at errors in the data processing, including harmonisation, these emerge while transforming responses into the final data set to be used for analysis. Processing errors seem to be of great importance: previous studies have repeatedly reported errors when classifying the country-specific educational qualifications into ISCED (Kieffer 2010; Ortmanns and Schneider 2016a; Schneider 2009; Hoffmeyer-Zlotnik 2008). Those errors directly influence the education distributions. We distinguish two survey characteristics here. The first one indicates *whether the official ISCED mapping is applied.* This is important because only if the educational qualifications are classified to ISCED in a consistent way, for example by following the official mappings, the education distributions are comparable across surveys (Schneider 2009). This characteristic distinguishes whether the assignment of ISCED codes to national education categories follows the official mapping or whether we find deviations from the official mapping. The EU-LFS and EU-SILC are conducted by the national statistical offices, which are also often responsible for developing countries' ISCED mapping, meaning they determine the ISCED code for each country-specific educational qualification. Therefore, we expect that

the EU-LFS and the EU-SILC follow the official mapping and that processing errors are rare in these surveys. In the other surveys, classification errors may occur more often because of lack of expertise in implementing the ISCED classification, which might lead to 'accidental' errors. The other reason for this processing error is lack of trust in the official mappings and this might lead to intended deviations from the official ISCED mapping. This deviation is more common in academic surveys such as ESS, EVS and ISSP, which are not obliged to follow the official ISCED mappings. Therefore, we estimate that differences in the application of the official ISCED mappings across surveys can contribute to inconsistencies in the education distribution (Hypothesis 10).

The second survey characteristic indicating processing or harmonisation error describes the *degree of centralisation when applying the ISCED classification* for the survey. It distinguishes between decentralised, partly centralised and centralised processing. In the decentralised approach, the country teams, who are familiar with their education system, are responsible for assigning the ISCED codes to national education categories. The EU-LFS and most other surveys implemented this approach. In contrast, in the centralised approach, one institute is responsible for assigning the ISCED codes for all countries of the survey. The Eurobarometer follows this method. Applying ISCED codes for several countries requires much expertise in ISCED and in the different educational systems. If one of these components is lacking, the chance of processing or harmonisation errors increases. Another approach combines both methods: classifying the national education category in ISCED is carried out by the country teams, but it is also checked centrally. This is beneficial because it involves country experts and an expert in the application of ISCED, and aims to optimise cross-national comparability. The ESS implemented this approach. Hence, differences in the degrees of centralisation across the surveys can increase inconsistencies in the education distributions across surveys within the same countries and years (Hypothesis 11).

3.3. Survey Characteristics Related to Both Measurement and Representation

Two survey characteristics are related to both dimensions of the TSE framework: mode of data collection and fieldwork organisation. Regarding the *mode of data collection*, we distinguish between face-to-face interviews, telephone interviews, self-administered modes (including web and postal surveys), and mixed-mode designs. The mode is a relevant factor for representation because different modes tend to systematically over- or under-represent certain groups, for example web surveys tend to over-represent more highly educated respondents (Couper 2000; Dever et al. 2008). Regarding the measurement dimension, the mode indicates the presence of an interviewer and the communication channel used. In face-to-face or telephone interviews, the presence of an interviewer makes socially desirable answering and interviewer effects more likely (De Leeuw and Van der Zouwen 2001; Lyberg and Kasprzyk 2011), however, interviewers may also help the respondent identify a suitable answer. In face-to-face or self-administered modes, respondents usually see a list of education categories, while in telephone interviews, these categories are read out or an open response is coded by the interviewer, which is more error-prone and primacy or recency effects can occur in the former case (Noelle-Neumann and Petersen 2000). Therefore, we expect that different

modes of data collection across the surveys within the same countries and years can increase inconsistencies in the education distributions across surveys (Hypothesis 12).

Fieldwork agencies are responsible for conducting the survey and are thereby involved in several aspects of sample representation and measurement. Therefore, the fieldwork agency can be seen as indicator for the standard of the survey and as proxy for different aspects, including those that could not be specified as survey characteristic due to a lack of information. This, for instance, applies to the availability of information on interviewer training. Concerning the EU-LFS, we would expect the overall standard to be quite high, largely because the fieldwork is done by a public authority, mostly the national statistical offices. This also applies to the second official survey, the EU-SILC. For the other surveys, commonly other fieldwork agencies are responsible, e.g., universities, other scientific or commercial institutes. We hypothesise that different kinds of fieldwork agencies can contribute to inconsistencies in the education distributions across surveys within the same countries and years (Hypothesis 13).

4. Data, Variables and Methods

In this study, we analyse the impact of surveys' methodological characteristics on discrepancies between the distributions of the harmonised education variable when comparing the EU-LFS with nine other surveys within the same countries and years. A description of the EU-LFS and the other surveys was already given in Subsection 2.1. This study focuses on these surveys from the period 2008 to 2012. If a survey was run several times during this time, such as the EU-SILC, the Eurobarometer, the ESS and the ISSP, it is only included once in order not to overestimate its effect. For most surveys the education distribution is stable over the years, as long as the country-specific measurement instruments and the harmonised education variable do not change (Ortmanns and Schneider 2016b, 2016a). When deciding which year to include, we consider the following factors: (a) number of countries covered, (b) completeness of documentation of survey characteristics, (c) whether its harmonised education variable has systematically changed (as in the ESS 2010 and the ISSP 2011), in which case the most recent year is included, (d) when a single country is not present in the selected year, information from an earlier round is used for this country. Due to a consequential processing error in the ISCED variable for Iceland in the EU-LFS 2011 and 2012 (for details see Ortmanns and Schneider 2016a), data before 2011 are included as far as possible. Thus, we include the EU-SILC and the Eurobarometer of 2010, and the ESS and ISSP of 2012.

As described in Subsection 2.2, the dependent variable is Duncan's Dissimilarity Index that compares the education distributions for each country and year of the EU-LFS with the respective country and year of each other survey. The independent variables reflect the survey characteristics (see Section 3) that differ across surveys for the same country-year comparison. Annex 3 (Supplemental material) provides basic descriptions of each survey characteristic. As mentioned, we focus on whether the survey characteristics differ between the EU-LFS and the respective other survey. Thus, most variables are coded as binary and distinguish whether the survey characteristics are 'equal' (0) or 'unequal' (1). The variables on response categories, fieldwork duration, response rates, sample size and the index of gender and age distribution are operationalised in a slightly more nuanced

way. As described in Subsection 3.2, we generate an index to assess the comparability of the response categories and distinguish between equal, similar and different. When comparing the fieldwork duration of the EU-LFS with the other surveys, we distinguish between the following categories: 'equal fieldwork duration to the EU-LFS', including up to five percentage points more or fewer days than the EU-LFS, 'longer duration: up to 90 days' and 'longer duration: 90 days or more', 'shorter duration: up to 90 days'. These four categories cover all comparisons. Regarding response rates, we use the ones reported in the survey documentation, even when we do not know exactly how these have been calculated, which may hamper their comparability. For the comparison of the response rates, we generate the following categories: 'equal response rate to the EU-LFS' if the response rate is up to 5 percentage points lower or higher than in the EU-LFS, 'lower response rate: up to 30 percentage points, 'lower response rate: 30 percentage points or more' and 'higher response rate: up to 30 percentage points'. A category indicating a higher response rate of more than 30 percentage points was not required. Unfortunately, the Eurobarometer does not provide information on response rates and for some countries of the other surveys the response rates are not documented. In order to be able to include those anyway, we generate an additional category 'information not available'. The categories of the variables on fieldwork duration and response rate are based on their distributions, and in order to avoid small or empty categories, they are rather broad. We include these categories as dummy variables in the analysis, and the categories indicating equal response rate or fieldwork duration are used as reference categories. When comparing the sample sizes of the EU-LFS with the other surveys, we calculate the absolute differences in the sample size and then divide by 1,000 because of the very high number of respondents in the EU-LFS. We then include this as a continuous variable. Duncan's index on the gender and age distribution delivers percentages and these are directly included in the regression models.

For many of the survey characteristics analysed, it would be desirable to use a higher level of detail. Unfortunately, this is not possible due to large variation in the accessibility of information, and especially the quality and the richness of the documentation. Still we had to exclude single countries in single surveys from the analysis when the information on a survey characteristic was not available. Thereby the data set is reduced from 248 to 229 survey comparisons and their respective comparisons of survey characteristics. The highest number of countries covered for one comparison is 29 when comparing EU-LFS with the Eurobarometer, or the EQLS or the EWCS, whereas the comparison between EU-LFS and PIAAC contains only 12 countries. An overview of the countries participating in the surveys and those included in the analysis can be found in Table S4 (Supplemental material).

Survey characteristics may correlate with each other and also with the survey programmes. Multicollinearity could make it hard to properly disentangle the effects of individual variables. Therefore, we checked the correlations between the different survey characteristics beforehand and Cramer's V was below 0.65. More details can be found in the Tables showing cross tabulations and correlations for selected survey characteristics in Annex 4 (Supplemental material). Additionally, we calculate the Variance Inflation Factor (VIF) after each regression model.

In the analysis, we estimate four multiple OLS regression models to explore the impact of different survey characteristics on inconsistencies in the education distributions. The

first model shows the impact of the survey programmes alone and thereby illustrates the large variation in the education distributions across surveys. The survey comparisons are included as dummy variables, and the comparison of EU-SILC and EU-LFS is used as reference. To explain these inconsistencies through differences in the survey characteristics, the second model adds the survey characteristics related to representation of the population. The third model includes survey characteristics related to measurement and survey programmes. To further reduce multicollinearity we calculate the final model excluding the dummy variables of the survey programmes. This model focuses on the survey characteristics that show statistically significant effects in Models 2 and 3.

5. Results

5.1. Impact of the Survey Programmes

As seen in the boxplot diagram (see Figure 1, Subsection 2.2) the inconsistencies in the education distributions differ strongly across surveys within the same countries and years. As expected, this pattern recurs when running a linear regression to predict Duncan's Dissimilarity Index by the survey programmes alone.

Model 1 in Table 2 shows low values for the regression coefficients for PIAAC ($b = 2.30$) and the AES ($b = 2.38$) and these survey comparisons are not statistically significant. The regression coefficients of the comparisons to the other survey programmes are higher ($b > 5.00$) indicating larger inconsistencies in the education distribution than in the reference comparisons of EU-LFS and EU-SILC. The comparison of the EU-LFS and the ESS is significant at the five percent level ($p < .05$), and the comparisons of the EU-LFS to the Eurobarometer, the EQLS, the EWCS, the EVS and the ISSP are highly significant ($p < .001$).

The adjusted R^2 of this model is 17%, meaning 17% of the variance can be explained by just the surveys themselves. This is unexpected because we can imagine the survey programmes as 'containers' for different survey characteristics. To identify which survey characteristics contribute to the inconsistencies in the education distributions, we estimate further regression models.

5.2. Impact of Survey Characteristics Related to the Representation of the Population

In addition to the first model, this model (Model 2 in Table 2) includes the survey characteristics related to the representation of the population, namely: sampling design, final sampling unit, sample size, mandatory survey participation, fieldwork duration, response rate, Sodeur's Index and Duncan's Dissimilarity Index for the age and gender distributions. Mode of data collection and fieldwork agency are also included.

This model shows that adding variables related to representation does not improve model fit: The adjusted R^2 of this model is also 17%. To estimate the quality of this model relative to the first model, we calculate the Akaike Information Criterion (AIC). For Model 1, the AIC is 1650.8 and for this model the AIC slightly increases to 1664.4. The model that shows the lowest value of the AIC, here Model 1, performs best. Regarding multicollinearity, the highest value of the VIF in this model is 7.1, which we observe for the dummy variable of the Eurobarometer. This indicates that the

Table 2. Results from regression analyses estimating the impact of survey characteristics on the inconsistencies in the education distribution across surveys.

	Model 1			Model 2			Model 3			Model 4		
Predictor	b	SE	p	b	SE	p	b	SE	p	b	SE	p
Survey:(ref: SILC)												
AES	2.38	2.47	.337	4.25	3.10	.172	−1.05	3.06	.731			
PIAAC	2.30	3.04	.451	4.42	3.72	.237	−3.08	4.21	.464			
EB	12.97***	2.36	<.001	14.94**	4.63	.001	3.71	7.19	.606			
EQLS	8.33***	2.36	<.001	11.70**	4.01	.004	−0.70	4.34	.872			
EWCS	10.21***	2.36	<.001	14.22**	4.06	.001	1.40	4.30	.745			
ESS	5.94*	2.38	.013	8.55*	3.62	.019	−2.74	6.78	.686			
EVS	9.20***	2.38	<.001	12.34**	3.69	.001	1.59	4.22	.708			
ISSP	11.57***	2.47	<.001	13.95***	3.88	<.001	0.50	4.36	.909			
Different sampling design (ref: equal)				3.67*	1.52	.016	3.37**	1.25	.007	3.43**	1.25	.007
Different sampling unit (ref: equal)				0.51	1.75	.772						
Differences in Sodeur's index (ref: equal)				−1.11	2.57	.665						
Duncan's index age/gender				−0.10	0.19	.594						
Sample size/1000				0.00	0.01	.647						
Fieldwork duration: (ref: equal)												
Shorter, < 90 days				1.09	2.28	.631						
Longer, < 90 days				0.16	2.22	.941						
Longer, ≥ 90 days				−0.73	2.53	.772						
Response rate: (ref: equal)												
Higher, < 30 percentage points				−0.90	3.11	.772						
Lower, < 30 percentage points				−3.55	2.66	.183						
Lower, ≥ 30 percentage points				−5.14	2.93	.081						
Not available				−4.01	3.91	.307						
Differences in mandatory participation (ref: equal)				0.82	1.32	.537						
Different mode (ref: equal)				2.15	1.43	.134	1.44	1.19	.225	1.46	1.19	.217
Different agency (ref: equal)				−1.42	2.42	.557	−0.31	2.17	.886	1.34	1.71	.432
Differences in proxy-reporting (ref: equal)							3.97	3.21	.217	2.89	1.98	.148
Difference in using register (ref: equal)							−0.65	1.93	.735	−0.47	1.92	.806

Table 2. Continued.

Predictor	Model 1			Model 2			Model 3			Model 4		
	b	SE	p	b	SE	p	b	SE	p	b	SE	p
Response categories: (ref: equal)												
Similar							0.87	2.39	.714	0.67	2.37	.776
Different							5.13*	2.33	.029	5.28*	2.24	.020
Differences in centralised coding (ref: equal)							0.53	5.67	.926	0.86	1.25	.494
Differences in ISCED coding (ref: equal)							9.20***	1.31	<.001	9.39***	1.27	<.001
Constant	5.58**	1.71	.001	5.63	3.62	.122	2.15	2.27	.346	1.84	2.18	.399
Adjusted R^2 (%)	16.61			16.67			35.59			34.00		
Akaike information criterion (AIC)	1650.76			1664.42			1600.04			1598.15		
Mean of variance inflation factor (VIF)	1.75			3.43			7.17			1.79		
Number of observations	229			229			229			229		

Notes: Data sources, see in online Supplemental material.

Eurobarometer correlates with the analysed survey characteristics. The mean value of the VIF of this model is 3.4, which is higher than in Model 1 (mean VIF of 1.8) but still unproblematic.

The only survey characteristic that has a statistically significant impact ($p < 0.05$) in this model is different sampling designs across the surveys. The regression coefficient of 3.7 indicates that different sampling designs increase the inconsistencies in the education distributions by roughly four percentage points, compared with equal designs. Thus, we do not reject hypothesis H1. From the results of this model, we find no evidence that the survey characteristics contribute to a higher inconsistency of the education distribution and therefore we do not reject H2 and H3 and we reject hypotheses H4 to H6, H12 and H13. In contrast to most survey characteristics, the survey effects remain significant and their regression coefficients even increase. Overall, this model shows that even when controlling for a substantial number of survey characteristics related to the representation of the population, the survey programmes themselves have by far the largest impact on the observed inconsistencies in the education distributions across surveys.

5.3. Impact of Survey Characteristics Related to Measurement

The third regression model shown in Table 2 focuses on the survey characteristics related to measurement. The following survey characteristics are included in this model: different response categories of the education question, proxy reporting, use of register information, applying of the official ISCED mappings and the degree of centralisation when applying ISCED. Also included are mode of data collection and fieldwork agency, which refer to both dimensions of the TSE, as well as the sampling design, which was significant in the second model. This model also controls for the survey programmes again.

This model has an adjusted R^2 of 36%, meaning more than one-third of the variance can now be explained. This is an increase of 19 percentage points compared to the previous models. The increase of the adjusted R^2 indicates a strong impact of survey characteristics related to measurement, over and above the effects of the surveys themselves. Compared to Models 1 and 2, the AIC decreases to 1600.0, which indicates a higher quality of this model. Concerning multicollinearity, the mean value of the VIF is 7.2, which is higher than in Models 1 and 2. In detail, we find high VIF values of around 20 for the dummy variables of the survey programmes for the Eurobarometer and the ESS, as well as the survey characteristic on the degree of centralisation when applying ISCED. This is not surprising because we know that this survey characteristic is strongly associated with the survey programme.

In this model, three survey characteristics have a statistically significant impact: different sampling designs, different response categories in the education item(s) and application of the official ISCED mapping. We find the strongest impact from the survey characteristic that indicates differences in whether the official ISCED mappings were applied between the EU-LFS and the surveys in question. This variable shows a high regression coefficient of 9.2, meaning inconsistency in the mapping of the national educational qualification into ISCED increases inconsistencies in the education distributions by roughly ten percentage points compared to consistent mapping. This effect is highly significant ($p < 0.001$). Thus, whether the official ISCED mappings are

applied is a crucial factor that explains deviations in the education distributions across surveys within countries and years. Therefore, we do not reject Hypothesis H10.

The survey characteristic indicating different response categories in the education items between the EU-LFS and the other surveys is also significant ($p < 0.05$). The regression coefficient of 5.1 indicates that using different response categories raises inconsistencies in the education distribution across surveys by roughly five percentage points compared to equal response categories. Thus, we also do not reject Hypothesis H7.

The survey characteristic assessing different sampling designs between the EU-LFS and other surveys, which was the only significant factor in Model 2, is again significant. The regression coefficient increases to 3.4 and the p-value is smaller in this model ($p < 0.01$), thus we again do not reject Hypothesis H1 in this model. Nevertheless, the effect of sampling design is smaller compared to the coefficients related to measurement.

All other survey characteristics are not statistically significant. The survey comparisons themselves are also not significant any more. Thus, in this model we identified the survey characteristics causing inconsistencies in the education distributions across surveys, and we successfully opened 'the black box of the surveys'.

In the final model (Model 4 in Table 2) the adjusted R^2 slightly decreases to 34%. The AIC declines to 1598.2, which is lowest value across all models, indicating that this is the best model estimated. Though excluding the survey programmes, we also reduce multicollinearity and the mean value of the VIF decreases to 1.8. The statistical significance of the variables assessing different sampling designs ($p < 0.01$), different response categories ($p < 0.05$) and differences in the application of the official ISCED mapping ($p < 0.001$) between the EU-LFS and the other surveys remain. This highlights the importance of these three survey characteristics independently of the survey programmes. Thus, we do not reject Hypotheses H1, H2, H3, H7 and H10, but according to this analysis, we can reject all other hypotheses. This result emphasises a predominant effect of measurement, especially the consistency of applying the official ISCED mappings and consistent response categories in the education question. Those are the key elements when it comes to explaining the inconsistencies in the education distributions across surveys within countries and years.

6. Conclusion and Discussion

This article asked which survey characteristics could explain the inconsistencies in the education distributions when comparing nine cross-national surveys to the EU-LFS. To answer that question, the impact of 15 survey characteristics and the survey programmes themselves were estimated. The data set used for this analysis contains detailed macro-information concerning the survey characteristics for the countries and years of the ten surveys. The main finding of this study is that differences in applying the official ISCED mappings (H10), differences in the response categories of the education question across surveys (H7), as well as – but to a lesser degree – differences in the sampling designs of the surveys (H1), are systematically related to inconsistencies in the education distributions across surveys within the same countries and years. These results are in line with our expectation and also with previous research (Kieffer 2010; Schneider 2009; Ortmanns and Schneider 2016a, 2016b) that focused on the measurement of the education

variable to explain inconsistent education distributions. Hence, the focus of previous studies was well justified. The comprehensive analysis of survey characteristics in this study additionally shows that apart from the sampling design, the survey characteristics related to the representation of the population do not cause inconsistencies in the education distribution across surveys.

To achieve higher consistency in the education distributions across surveys, survey organisers should, firstly, reduce the processing error by improving the assignment of the response categories of the education item to the ISCED classification. To make recommendations on how to reduce the processing error, we further need to distinguish whether the deviation from the official ISCED mapping occurs accidentally or whether it is intended. 'Accidental' errors, which are often caused by limited knowledge when assigning the national educational qualification to the ISCED classification, can be avoided through implementing additional quality checks and the application of the official ISCED mappings in principal (Ortmanns and Schneider 2016a).

In contrast, the intended deviations applied by some academic surveys aim to enhance comparability of cross-national education data across countries (Ortmanns and Schneider 2016a). This is justified because during the development and the implementation of the ISCED mappings it is vulnerable to political influence of education ministries and national statistical offices. The latter often develop the national ISCED mappings and they do not equally strictly apply the ISCED criteria. At the same time, some criteria formulated in the ISCED classification are rather vague and thus leave some room for interpretation. This explains why countries with similar qualification nevertheless classify them to different ISCED codes. The intended deviations made by academic surveys attempt to correct for this. However, these deviations also introduce incomparability across survey, notably with official surveys applying the official ISCED mappings, such as the EU-LFS and the EU-SILC. Intended deviations could be avoided when the quality control of the national ISCED mappings, for example through UNESCO, would become stricter. As this is currently not ensured, the international survey community has good reasons to find solutions to produce comparable education data for their own purpose. Academic surveys, for instance, could agree on applying an 'alternative' ISCED scheme that adjusts the official mappings to optimise comparability over time and space. This alternative version should be well-documented and contain recodes to the official mappings in order to still compare them with official education data.

The second important recommendation to achieve higher consistency in the education distributions across surveys is to improve the education item itself. We should aim for standardised country-specific education categories, which use a terminology that is equally understandable for everyone and avoid generic terms and descriptions. These categories can then be implemented in all surveys, national as well as international, that measure education as a background variable. Of course, no instrument will be without measurement error; however, if every survey uses the same instrument, the error will be consistent and this enhances data comparability. The development of these country-specific education categories and their assignment to ISCED should be done by a national expert group, which should consist of experts of the country-specific educational system, experts of ISCED and also representatives of the national statistical office, the education ministry as well as a survey expert. Ideally, also an expert in cross-national surveys should

be included in the discussion to consider comparability in international surveys. Additionally, for countries having a similar educational system, for instance Germany, Austria and Switzerland or the UK and Ireland, it is also worthwhile to exchange their suggestions and, even better, to discuss shared issues. Then we can also better consider comparability across *countries*, which we did not look at in this article.

This study also faces some limitations. An obvious one is the small number of cases (n = 229), which might be problematic for testing such a large number of survey characteristics. However, focusing on whether the survey characteristics are equal or unequal across surveys prevents us from having small or even empty cells. The disadvantage of these variables is that they are quite generic, and it is not possible to, for instance, to identify which kind of fieldwork agency (public authority including statistical office, university or other scientific institute, commercial institute) causes more or less inconsistent education distributions. We can only tell whether differences in the fieldwork agencies between the survey in question and the EU-LFS affect deviations in the education distribution. This structure of the variables and the low case number furthermore do not allow calculation of more complex models or application of multilevel modeling.

Another limitation of this study is that it compares the education distribution using the 1997 version of ISCED, whereas surveys are increasingly implementing the more recent version – ISCED 2011. However, we are convinced that the current results would not be very different and we would still find inconsistencies when comparing the education distributions across surveys within countries and years. One change in ISCED 11 is a better differentiation of levels within tertiary education, so when surveys implement this new version, they will be paying particular attention to the codes of tertiary education. However, we observe the greatest inconsistencies for ISCED level 3 (upper secondary education), and also find deviations in the adjacent categories ISCED level 2 (lower secondary) and ISCED level 4 (post-secondary, non-tertiary). At these levels we find most of the ambiguous terms and generic descriptions used in the response categories of the surveys, especially with the vocational qualifications. These can also cause errors when assigning ISCED codes. The inconsistencies on these levels will not disappear when implementing ISCED 11, unless surveys start primarily to correct for accidental errors when assigning ISCED codes and update the country-specific response categories alongside the implementation of the new ISCED version. The ESS in 2010 undertook such a detailed check and updated its variables, and a similar review took place for the EVS 2017. The ISSP is currently considering how best to implement ISCED 11. The effort invested in the education variables in these surveys is likely to reduce inconsistencies in the education distribution in the future.

An output of this study is the data file of survey characteristics that is publicly available at the SowiDataNet|datorium (Ortmanns 2020). Until recently, survey characteristics have rarely been considered in substantive data analyses, and only few studies exist that include them (e.g., Heath et al. 2009; Van Tuyckom and Bracke 2014). The main reason that survey characteristics are often neglected is probably that collecting and harmonising this information requires considerable effort. Often the documentation of survey characteristics is neglected, meaning we have to look at several documents of varying quality, to be found on different webpages of the surveys or data archives. Sometimes we still cannot find complete information, and it is little standardised. More systematic and

easily accessible documentation would be very helpful. This would enhance transparency and increase the possibility of developing standards on how to report survey characteristics. Some initiatives have begun by collecting, documenting and publishing information on methodological survey characteristics relevant for their specific projects. Such an initiative exists for official statistics within the online platform MISSY, which provides metadata of the EU-LFS and EU-SILC. A further initiative that recently has been completed is part of the EU project 'Synergies for Europe's Research Infrastructures in the Social Sciences'. In work package two, the sampling practices of European surveys have been documented to compare and finally improve them (Scherpenzeel et al. 2017). The ongoing research project on survey data harmonisation of the Polish Academy of Sciences in cooperation with Ohio State University also devotes substantial effort to documenting and harmonising data related to democratic values and protest behaviours (Słomczyński et al. 2018). Unfortunately, this study was already underway, so the outcomes of these initiatives could only be used for cross-checking. Finally, the IPUMS-International project, a collaboration of the University of Minnesota, National Statistical Offices, international data archives, as well as other international organisations, harmonises publicly available census data and provides a systematic inventory (Minnesota Population Center 2019). Unfortunately, it does not (yet) offer a harmonised ISCED variable that can be used for cross-national comparisons. However, all these projects will facilitate future studies like this, as well as substantive (rather than methodological) studies that would like to control for the impact of a single survey characteristic.

7. References

Becker, G.S. 1993. *Human Capital. A Theoretical and Empirical Analysis with Special Reference to Education*, 3rd ed. Chicago: The University of Chicago Press.

Bekhuis, H., Lubbers, M., Verkuyten. M. 2014. "How Education Moderates the Relation Between Globalization and Nationalist Attitudes." *International Journal of Public Opinion Research* 26(4): 487–500. DOI: https://doi.org/10.1093/ijpor/edt037.

Bethlehem, J.D., F. Cobben, and B. Schouten. 2011. *Handbook of Nonresponse in Household Surveys*. Hoboken: John Wiley & Sons.

Biemer, P.P. and L.E. Lyberg. 2003. *Introduction to Survey Quality*. Hoboken: John Wiley & Sons.

Blair, J., G. Menon, and B. Bickart. 2011. "Measurement Effects in Self vs. Proxy Response to Survey Questions: An Information-Processing Perspective." In *Measurement Errors in Surveys*, edited by P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Matthiowetz, S. Sudman, 145–166. Hoboken: John Wiley & Sons.

Blau, P.M. and O.D. Duncan. 1967. *The American Occupational Structure*. New York: John Wiley & Sons.

Bol, T. and H.G. Van de Werfhorst. 2013. "Educational Systems and the Trade-Off Between Labor Market Allocation and Equality of Educational Opportunity." *Comparative Education Review* 57(2): 285–308.

Bolzendahl C.I. and D.J. Myers. 2004. "Feminist Attitudes and Support for Gender Equality: Opinion Change in Women and Men, 1974–1998." *Social Forces* 83(2): 759–789. DOI: https://doi.org/10.1353/sof.2005.0005.

Braun, M. and P. Ph. Mohler. 2003. "Background Variables." In *Cross-Cultural Survey Methods*, edited by J.A. Harkness, F.J.R. Van de Vijver, and P. Ph. Mohler, 101–115. Hoboken: John Wiley & Sons.

Breen, R. and J.O. Jonsson. 2000. "Analyzing Educational Careers: A Multinomial Transition Model." *American Sociological Review* 65(5): 754–772. DOI: https://doi.org/10.2307/2657545.

Breen, R. and J.O. Jonsson. 2005. "Inequality of Opportunity in Comparative Perspective: Recent Research on Educational Attainment and Social Mobility." *Annual Review of Sociology* 31: 223–243. DOI: https://doi.org/10.1146/annurev.soc.31.041304.122232.

Church, A.H. 1993. "Estimating the Effect of Incentives on Mail Survey Response Rates: A Meta-Analysis." *Public Opinion Quarterly* 57(1): 62–79. DOI: https://doi.org/10.1086/269355.

Coenders, M. and P. Scheepers. 2003. "The Effect of Education on Nationalism and Ethnic Exclusionism: An International Comparison." *Political Psychology* 24(2): 313–343. DOI: https://doi.org/10.1111/0162-895X.00330.

Couper, M.P. 2000. "Web Surveys: A Review of Issues and Approaches." *Public Opinion Quarterly* 64(4): 464–494. DOI: https://doi.org/10.1086/318641.

Daikeler, J., M. Bosnjak, and K. Lozar-Manfreda. 2019. "Web Versus Other Survey Modes: An Updated and Extended Meta-Analysis Comparing Response Rates." *Journal of Survey Statistics and Methodology*. DOI: https://doi.org/10.1093/jssam/smz008.

De Leeuw, E.D. and J. Van der Zouwen. 2001. "Data Quality in Telephone and Face-to-Face Surveys: A Comparative Meta-Analysis." In *Telephone Survey Methodology*, edited by R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls II, and J. Waksberg, 283–300. New York: John Wiley & Sons.

Dever, J.A., A. Rafferty, and R. Valliant. 2008. "Internet Surveys: Can Statistical Adjustments Eliminate Coverage Bias?" *Survey Research Methods* 2(2): 47–60. DOI: https://doi.org/10.18148/srm/2008.v2i2.128.

Duncan, O.D. and B. Duncan. 1955. "A Methodological Analysis of Segregation Indexes." *American Sociological Review* 20(2): 210–217. DOI: https://doi.org/10.2307/2088328.

Ehling, M. 2003. "Harmonising Data in Official Statistics: Development, Procedures, and Data Quality." In *Advances in Cross-National Comparison. A European Working Book of Demographic and Socio-Economic Variables*, edited by J.H.P. Hoffmeyer-Zlotnik and C. Wolf, 17–31. New York: Kluwer Academic/ Plenum Publishers.

ESS. 2014a. *Documentation of ESS Post-Stratification Weights*. Available at: https://www.europeansocialsurvey.org/docs/methodology/ESS1-5_post_stratification_weights_documentation.pdf (accessed December 2019).

ESS. 2014b. *Weighting European Social Survey Data*. Available at: https://www.europeansocialsurvey.org/docs/methodology/ESS_weighting_data_1.pdf (accessed December 2019).

ESS. 2016a. European Social Survey Round 4 Data 2008. Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC data file version 4.4.

ESS. 2016b. European Social Survey Round 5 Data 2010. Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC data file version 3.3.

ESS. 2016c. European Social Survey Round 6 Data 2012. Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC data file version 2.3.

Eurofound. 2011. *European Working Conditions Survey (EWCS) 2010*. Colchester: UK Data Archive. Study number 6971 data file version 1. DOI: https://doi.org/10.5255/UKDA-SN-6971-1.

Eurofound. 2014. *European Quality of Life Survey (EQLS) 2011–2012*. Colchester: UK Data Archive. Study number 7316 data file version 3. DOI: https://doi.org/10.5255/UKDA-SN-7316-2.

European Commission. 2012. Eurobarometer 73.2 & 73.3 (2-32010). GESIS Data Archive, Cologne. ZA5236 data file version 2.0.1. DOI: https://doi.org/10.4232/1.11473.

Eurostat. 2008. European Union Labour Force Survey (EU-LFS) 2008. User Database, data file version 2016.

Eurostat. 2010a. European Union Labour Force Survey (EU-LFS) 2010. User Database, data file version 2016.

Eurostat. 2010b. European Union Statistics on Income and Living Conditions (EU-SILC) 2010. User Database, data file version CROSS-2010-6.

Eurostat. 2011a. Adult Education Survey (AES). User Database, data file version 1.

Eurostat. 2011b. European Union Labour Force Survey (EU-LFS) 2011. User Database, data file version 2016.

Eurostat. 2012. European Union Labour Force Survey (EU-LFS) 2012. User Database, data file version 2016.

Eurostat. 2019. Smarter, Greener, More Inclusive? Indicators to Support The Europe 2020 Strategy. Luxembourg: Publication Office of the European Union. Available at: https://ec.europa.eu/eurostat/documents/3217494/10155585/KS-04-19-559-EN-N.pdf/b8528d01-4f4f-9c1e-4cd4-86c2328559de (accessed December 2019).

EVS. 2016. European Values Study 2008, 4th wave. Integrated Dataset. GESIS Data Archive, Cologne. ZA4800 data file version 4.0.0. DOI:https://doi.org/10.4232/1.12458.

Groves, R.M. and M.P. Couper. 1998. *Nonresponse in Household Interview Surveys*. New York: John Wiley & Sons.

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.

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

Groves, R.M. and L.E. Lyberg. 2010. "Total Survey Error: Past, Present, and Future." *Public Opinion Quarterly* 74(5): 849–879. DOI: https://doi.org/10.1093/poq/nfq065.

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.

Heath, A., J. Martin, and T. Spreckelsen. 2009. "Cross-National Comparability of Survey Attitude Measures." *International Journal of Public Opinion Research* 21(3): 293–315. DOI: https://doi.org/10.1093/ijpor/edp034.

Hoffmeyer-Zlotnik, J.H.P. 2008. "Harmonisation of Demographic and Socio-Economic Variables in Cross-National Survey Research." *Bulletin de Méthodologie Sociologique* 98: 5–24. DOI: https://doi.org/10.1177/075910630809800103.

Hyman, H.H. and C.R. Wright. 1979. *Education's Lasting Influence on Values*. Chicago: University of Chicago Press.

ISSP Research Group. 2015. International Social Survey Programme: Health and Health Care - ISSP 2011. GESIS Data Archive, Cologne. ZA5800 data file version 3.0.0. DOI: https://doi.org/10.4232/1.12252.

ISSP Research Group. 2016. International Social Survey Programme: Family and Changing Gender Roles IV - ISSP 2012. GESIS Data Archive, Cologne. ZA5900 data file version 4.0.0. DOI: https://doi.org/10.4232/1.12661.

Kalmijn, M. 2003. "Country Differences in Sex-Role Attittudes: Cultural and Economic Explanations." In *The Cultural Diversity of European Unity. Findings, Explanations and Reflections from the European Values Study*, edited by W. Arts, L.C.J.M. Halmann and J.A.P. Hagenaars: 311–337. Leiden: Koninklijke Brill.

Kaminska, O. and P. Lynn. 2017. "Survey-Based Cross-Country Comparisons Where Countries Vary in Sample Design: Issues and Solutions." *Journal of Official Statistics* 33(1): 123– 136. DOI: https://doi.org/10.1515/JOS-2017-0007.

Kieffer, A. 2010. "Measuring and Comparing Levels of Education: Methodological Problems in the Classification of Educational Levels in the European Social Surveys and the French Labor Force Surveys." *Bulletin de Méthodologie Sociologique* 107: 49–73. DOI: https://doi.org/10.1177/0759106310369974.

Kleven, Ø. and K. Ringdal. 2017. Level of Education – Measuring the Quality of Questions in Survey Interviews by Administrative Records on Education. Experiences from the Norwegian European Social Survey 2004 – 2014. In *Proceedings of the 7th Conference of the European Survey Research Association (ESRA)*, July 20, 2017. Lisbon. Available at: https://www.europeansurveyresearch.org/conference/programme2017?sess=210#582 (accessed April 2020).

Kohler, U. 2007. "Surveys from Inside: An Assessment of Unit Nonresponse Bias With Internal Criteria." *Survey Research Methods* 1(2): 55–67. DOI: https://doi.org/10.18148/srm/2007.v1i2.75.

Kohler, U. 2008. "Assessing the Quality of European Surveys - Towards an Open Method of Coordination for Survey Data." In *Handbook of Quality of Life in the Enlarged European Union*, edited by J. Albers, T. Fahey, and C. Saraceno, 405–423. London, New York: Routledge.

Kreuter, F., S. Eckman, K. Maaz, and R. Watermann. 2010. "Children's Reports of Parents' Education Level: Does It Matter Whom You Ask and What You Ask About?" *Survey Research Methods* 4(3): 127–138. DOI: https://doi.org/10.18148/srm/2010.v4i3.4283.

Lohr, S.L. 2009. *Sampling: Design and Analysis*. 2nd ed. Boston: Brooks/Cole Cengage Learning.

Lyberg, L.E. and D. Kasprzyk. 2011. "Data Collection Methods and Mesurement Error: An Overview." In *Measurement Errors in Surveys*, edited by P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowwetz, and S. Sudman, 235–257. Hoboken: John Wiley & Sons.

Minnesota Population Center. 2019. *Integrated Public Use Microdata Series, International*. Data Minneapolis, MN: IPUMS, data file version 7.2. DOI: https://doi.org/10.18128/D020.V7.2.

Moore, J.C. 1988. "Self/Proxy Response Status and Survey Response Quality: A Review of the Literature." *Journal of Official Statisics* 4(2): 155–172.

Müller, W. and W. Karle. 1993. "Social Selection in Educational Systems in Europe." *European Sociological Review* 9(1): 1–23. DOI: https://doi.org/10.1093/oxfordjournals.esr.a036652.

Noelle-Neumann, E. and T. Petersen. 2000. "Das halbe Instrument, Die halbe Reaktion. Zum Vergleich von Telefon- und Face-to-Face Umfragen." In *Methoden in Telefonumfragen*, edited by V. Hüfken, 183–200. Wiesbaden: Westdeutscher Verlag.

OECD. 2013. Programme for the International Assessment of Adult Competencies (PIAAC) 2011 - Public Use Files: data file version 1.0. Available at: https://www.oecd.org/skills/piaac/data/ (accessed April 2020).

OECD. 2015. *Education at a Glance 2015: OECD Indicators*. OECD Publishing. Available at: https://www.oecd-ilibrary.org/education/education-at-a-glance-2015_eag-2015-en (accessed December 2019).

OECD. 2016. *Education at a Glance 2016: OECD Indicators*. OECD Publishing. Available at: https://www.oecd-ilibrary.org/education/education-at-a-glance-2016_eag-2016-en (accessed December 2019).

OECD. 2017. *Education at a Glance 2017: OECD Indicators*. OECD Publishing. Available at: https://www.oecd-ilibrary.org/education/education-at-a-glance-2017_eag-2017-en (accessed December 2019).

OECD, and Eurostat. 2014. *Joint Eurostat-OECD Guidelines on the Measurement of Educational Attainment in Household Surveys*. Available at: http://ec.europa.eu/eurostat/documents/1978984/6037342/Guidelines-on-EA-final.pdf (accessed December 2019).

Ortmanns, V. and S.L. Schneider. 2016a. "Can We Assess Survey Representativeness of Cross-National Surveys Using the Education Variable?". *Survey Research Methods* 10(3): 189– 210. DOI: https://doi.org/10.18148/srm/2016.v10i3.6608.

Ortmanns, V. and S.L. Schneider. 2016b. "Harmonization Still Failing? Inconsistency of Education Variables in Cross-National Public Opinion Surveys." *International Journal of Public Opinion Research* 28(4): 562–582. DOI: https://doi.org/10.1093/ijpor/edv025.

Ortmanns. V. 2020. Education distributions and survey characteristics of ten cross-national surveys. GESIS, SowiDataNet|datorium. Data file Version: 1.0.0. DOI: https://doi.org/10.7802/1.2002.

Peytcheva, E. and R.M. Groves. 2009. "Using Variation in Response Rates of Demographic Subgroups as Evidence of Nonresponse Bias in Survey Estimates." *Journal of Official Statistics* 25(2): 193–201.

Schanze, J.-L. 2017. *Report on Sampling Practices for the Institutionalized Population in Social Surveys*. Deliverable 2.16 of the SERISS project funded under the European Union's Horizon 2020 research and innovation programme GA No: 654221. Available at: www.seriss.eu/resources/deliverables (accessed December 2019).

Scherpenzeel, A., A.M. Maineri, J. Bristle, S. Pflüger, I. Mindarova, S. Butt, S. Zins, T. Emery, and R. Luijkx. 2017. *Report on the Use of Sampling Frames in European Studies*. Deliverable 2.1 of the SERISS project funded under the European Union's Horizon 2020 research and innovation programme GA No: 654221. Available at: www.seriss.eu/resources/deliverables (accessed December 2019).

Schneider, S.L. 2009. *Confusing Credentials: The Cross-Nationally Comparable Measurement of Educational Attainment*. (PhD thesis). Oxford: University of Oxford, Nuffield College. Available at: http://ora.ouls.ox.ac.uk/objects/uuid:15c39d54-f896-425b-aaa8-93ba5bf03529 (accessed December 2019).

Schneider, S.L., D. Joye, and C. Wolf. 2016. "When Translation Is Not Enough: Background Variables in Comparative Surveys." In T*he SAGE Handbook of Survey Methodology*, edited by C. Wolf, D. Joye, T.W. Smith, and Y.-c. Fu, 288–307. London: Sage Publications.

Semyonov, M., R. Raijman, and A. Gorodzeisky. 2008. "Foreigners' Impact on European Societies Public Views and Perceptions in a Cross-National Comparative Perspective." *International Journal of Comparative Sociology* 49(1): 5–29. DOI: https://doi.org/10.1177/0020715207088585.

Singer, E. and Ye, C. 2013. "The Use and Effects of Incentives in Surveys." *The ANNALS of the American Academy of Political and Social Science* 645(1): 112–141. DOI: https://doi.org/10.1177/0002716212458082.

Słomczyński, K.M., Jenkins, J.C., Tomescu-Dubrow, I., Kołczyńska, M., Wysmułek, I., Oleksiyenko, O., Powałko, P. and Zielinski, M.W. 2018. "Survey Data Recycling Dataverse - Master Box." Harvard Dataverse, data file version 1.1. DOI: https://doi.org/10.7910/DVN/VWGF5Q.

Słomczyński, K.M., Tomescu-Dubrow, I., Jenkins, J.C., Kołczyńska, M., Powałko, P., Wysmułek, I., Oleksiyenko, O., Zielinski, M.W. and Dubrow, J.K. 2016. *Democratic Values and Potest Behavior - Harmonization of Data from International Survey Projects*. Warsaw: IFiS Publishers.

Smith, T.W. 1995. "Some Aspects of Measuring Education." *Social Science Research* 24(3): 215–242. DOI: https://doi.org/10.1006/ssre.1995.1008.

Smith, T.W. 2010. "The Globalization of Survey Research." In *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, edited by J.A. Harkness, M. Braun, B. Edwards, T.P. Johnson, L.E. Lyberg, P.Ph. Mohler, B.-E. Pennell, and T.W. Smith, 477–484. Hoboken: John Wiley & Sons.

Smith, T.W. 2011. "Refining the Total Survey Error Perspective." *International Journal of Public Opinion Research* 23(4): 464–484. DOI: https://doi.org/10.1093/ijpor/edq052.

Sodeur, W. 1997. "Interne Kriterien zur Beurteilung von Wahrscheinlichkeitsauswahlen." In *ZA-Information / Zentralarchiv Für Empirische Sozialforschung*, 41: 58–82. Cologne. Available at: https://www.ssoar.info/ssoar/handle/document/19999 (accessed December 2019).

Sodeur, W. 2007. "Entscheidungsspielräume von Interviewern bei der Wahrscheinlichkeitsauswahl - Ein Vergleich von ALLBUS-Erhebungen." *MDA - Methoden Daten Analysen* 1(2): 107– 130.

UNESCO-UIS. 2006. *International Standard Classification of Education ISCED 1997*. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000146967?posInSet=1&queryId=a8d996ff-eddd-4153-9468-5d7f9e8f7199 (accessed December 2019).

Van Tuyckom, C. and P.F. Bracke. 2014. "Survey Quality and Cross-National Sports Research: A Case Study of the 2007 ISSP Survey." *European Journal of Sport Science* 14(1): 228– 234. DOI: https://doi.org/10.1080/17461391.2012.683814.

Weakliem, D.L. 2002. "The Effects of Education on Political Opinions: An International Study." *International Journal of Public Opinion Research* 13(2): 141–157. DOI: https://doi.org/10.1093/ijpor/14.2.141.

Received August 2018
Revised February 2019
Accepted December 2019

sciendo

Journal of Official Statistics, Vol. 36, No. 2, 2020, pp. 411–434, http://dx.doi.org/10.2478/JOS-2020-0021

Investigating the Effects of the Household Budget Survey Redesign on Consumption and Inequality Estimates: the Italian Experience

Nicoletta Pannuzi[1], *Donatella Grassi*[1], *Achille Lemmi*[2], *Alessandra Masi*[1], *and Andrea Regoli*[3]

In 2014, many innovations were introduced in the Italian Household Budget Survey (HBS) in response to changes in European recommendations and purchasing behaviours and to an increased demand for information in the context of social and economic research. New instruments and techniques have been introduced, together with more accurate methodologies, with the aim of improving the survey, by both reducing the bias and variance of survey estimates and supplying estimation for additional subpopulations and variables. Given the parallel conduction of the former and new HBS in 2013, it has been possible to evaluate the effects of the abovementioned changes on consumption expenditure and inequality estimates and to compare the sample representativeness of selected subpopulations in both surveys.

Key words: Survey design; data quality; zero expenditures; post-stratification.

1. Introduction

Household Budget Surveys (HBS) are conducted in all EU member states and several other countries (OECD 2013); they mainly focus on consumption expenditure and have a primary aim (at the national level) of calculating weights for the Consumer Price Index. Since 1988, Eurostat has collected and disseminated these survey data every five years (Eurostat 2015).

In the most recent decades, a European (legally non-binding) agreement was made with the aim of developing shared definitions and methodologies to improve the quality and comparability of the HBSs. However, in the absence of a European framework regulation (Eurostat 2017, 2003), the national HBSs still differ in various aspects, ranging from methodologies to data collection techniques and definitions of variables.

[1] Italian National Institute of Statistics, Via Cesare Balbo, 16–00184 Rome, Italy. Emails: nicoletta.pannuzi@istat.it, donatella.grassi@istat.it, and alessandra.masi@istat.it
[2] ASEDS, Tuscan Interuniversity Research Centre "Camilo Dagum" University of Pisa, Department of Economics and Management, Via Cosimo Ridolfi, 10–56124 Pisa, Italy. Email: lemmiachille@virgilio.it
[3] University of Naples Parthenope, Department of Management and Quantitative Studies, Via Generale Parisi, 13–80132 Naples, Italy. Email: andrea.regoli@uniparthenope.it
Acknowledgments: The Authors would like to express their gratitude to the members of the interinstitutional working group on poverty estimate – established in 2015 by the Italian National Institute of Statistics – for their valuable and constructive suggestions during the research work. The opinions expressed in this article solely represent those of the Authors and do not necessarily reflect the official viewpoint of the Italian National Institute of Statistics.

In Italy, the survey has been conducted regularly since the 1960s; in 2014, it was completely redesigned (for details, see Grassi and Pannuzi 2015) to consider the new European recommendations and the changes in expenditure and consumption behaviours of the population. The purpose of the redesign was also to improve the survey, by reducing the variance of survey estimates and supplying estimation for additional subpopulations and variables.

Incorporating the European recommendations required change in both methodological issues and survey design relative to different aspects, including field of observation, sampling design and survey technique, and expenditure item classification.

The availability of new products in the market, the changes in the distribution channels and the important social, economic and cultural transformations of recent years (such as the increased foreign presence and changes in labour market participation) have modified the Italian dynamics of purchasing behaviours. This change has required new instruments, techniques and methods to properly collect consumption expenditure data.

All of these changes have entailed marked and statistically significant differences in many estimates produced by the survey (Istat 2015). Lemmi et al. (2019) analysed the differences between poverty estimates from the new and old surveys (hereafter new and former HBS), showing that the introduced innovations improved the estimation quality and determined a significant reduction in the number of households and individuals classified as poor.

In this article, we evaluate the effect of the redesign on the level, distribution and inequality of consumption expenditures and compare the sample coverage of selected subpopulations in both surveys. Specifically, we intend to show how the changes strive for variance reduction and demonstrate how the new design is successful in meeting this goal. In addition to the overall effect of the redesign, we also assess the impact of specific changes through an *ad hoc* simulation that treats the new HBS data as if they were observed according to the methodology of the former HBS. This investigation of the effects of the survey redesign was made possible by the parallel conducting of both surveys in the last two quarters of 2012 and all quarters of 2013.

Although the specific issues reported in this article relate to the Italian context and to 2013 data, it can be argued that this does not compromise its relevance or timeliness because the considered aspects have their own methodological relevance regardless of the country and data updating process. The applied approach can, in fact, represent a reference paradigm for evaluating the effects on estimates due to changes in different survey aspects.

The structure of the article is described as follows. Section 2 first summarises the main points of the HBS redesign process, highlighting the main differences between the former and new HBSs, and subsequently compares the coverage of selected populations and selected estimates between the two surveys. Section 3 analyses the effects of the innovations on the estimates of levels and inequality measures. Finally, the impact of the survey redesign on the estimates for population subgroups is presented in Section 4. Section 5 contains the conclusions.

2. Former and New Italian Household Budget Surveys

In this section, we compare the main characteristics of the former and new HBS. We also evaluate the coverage error in estimates for specific subgroups and the variance of the estimates of consumption expenditure levels and inequality.

2.1. Population and Sample Design Features

For reasons related to survey costs and organisational issues, the samples of the former and new HBS are concentrated in a limited number (approximately 500) of the over 8,000 Italian municipalities. The consequent choice is a two-stage sample design, in which the first-stage units are the municipalities and the second stage units, or final sampling units, are the households.

In the former survey the municipalities were stratified, within each region, by demographic size only. In the new survey, the municipality typology is also inserted as a stratification variable distinguishing between metropolitan area (municipalities with over 250,000 inhabitants); large municipalities (municipalities in the periphery of the metropolitan area and municipalities with 50,000 inhabitants and more) and small municipalities (municipalities with less than 50,000 inhabitants). In these municipalities, the shares of the population are about 17%, 29% and 54%, respectively. The stratification by municipality typology accounts for the different levels and styles of consumption expenditure associated with living in municipalities of different sizes and more or less close to metropolitan areas.

The choice of using a deeper stratification has the twofold aim of reducing the variance of the estimates and assuring sufficient data to produce estimates for specific subpopulations.

The HBS, in Italy, represents one of the primary sources for estimating the quarterly households' final consumption in the National Accounts System. The sampling design is therefore defined with reference to a generic quarter of the year and is identically replicated for the four quarters; moreover, a monthly stratification of the quarterly sample is also carried out. Consequently, the temporal dimension can be considered an additional stratification variable of the sample, which allows taking into account the highly seasonal nature of some types of expenditure.

Besides, the municipality typology and the month of participation are also considered in calculating the weights used to expand the sample to the population. They are obtained as the product of the following three factors: (1) basic coefficient (reciprocal of the inclusion probability); (2) correction factor for nonresponse (inverse of the response rate); (3) correction factor to match known population totals.

The post-stratification adjustments to the following control totals used in the former HBS have been preserved in the new survey:

- *Resident population by geographical area* (North, Centre, South and Islands), *sex and age groups* (0–14, 15–29, 30–44, 45–59, 60–74, 75 or more), in order to take into account the different levels and composition of expenditure characterising individuals of different age and sex, but also people living in the southern or northern part of the country. Traditionally about 46% of the total population live in the North, where households present higher levels and different typologies of expenditures than in the Centre and the South, where the share of the population is approximately 20% and 34%, respectively, and
- *Resident population and households by region*, following the NUTS2 Eurostat classification, in order to account for any differences in terms of administrative rules and services availability which may have impact on consumptions expenditures of citizens.

To these, others post-stratification adjustments have been added in the new HBS:

- *Resident population and households by geographical area and municipality typology* (metropolitan area, large municipalities, small municipalities), to be consistent with the sample stratification,
- *Foreign population by geographical area and sex* to account for the growing size of the foreign population in the total resident population – in 2014 it represented almost 8% of the resident population – and their different habits and consumption levels,
- *Population of 15 years and over by geographical area, employment status and position* (manager/white collar, blue collar, entrepreneur/freelancer, self-employed, looking for a job, retired from work, other), considering that the condition and the professional position impact on the income levels, and therefore on the expenditure, of the population; this total is derived from the Istat Labour Force survey, and
- *Population and households by geographical area and participation month* to be consistent with the sample stratification.

2.2. *Comparison of the Main Survey Characteristics*

In addition to the sample issues mentioned in the above sub-section, other HBS differences are summarised in Table 1. Both surveys are conducted using face to face interviews (just one interview in the former HBS, two interviews in the new HBS) and diaries (two diaries in the former HBS, one diary in the new HBS).

The adoption of the 2013 COICOP classification of variables (UN Statistical Commission 2018) was one of the Eurostat recommendations (Eurostat 2013) intended to achieve better comparability of measurement among the HBSs of different countries and with the consumer price indices classification. This change has resulted in an increase in expenditure items from 265 to 482 items.

Also part of the change was replacing the single questionnaire with two questionnaires (Zezza et al. 2017; Ngandu et al. 2016; Angrisani et al. 2015; Barrett et al. 2015; Bee et al. 2015; Smith et al. 2014; Andreski et al. 2014). The two new questionnaires are administered through computer-assisted personal interviews (CAPI) rather than through the paper-and-pencil interviews (PAPI) used previously. One of the two questionnaires collects information on socio-demographic characteristics of household members and housing characteristics (that are invariant in the short term); the other collects less frequent or exceptional expenses for a predefined list of items (with empty spaces to be filled for unlisted items). The self-administered diary, as in the former HBS, is a paper diary used to collect daily or high-frequency expenditures; also in this case the items are predefined and listed, with the opportunity of adding new ones. Instead of collection through a separate diary, as in the former survey, in the new HBS, self-consumption is collected in a dedicated section of the daily expenditure diary following the COICOP classification. The diaries are collected by an interviewer who converts them to electronic format using a computer-assisted input (CADI) system.

In both surveys, substitutions to replace nonresponding households are allowed (for details, see Freguja and Romano 2014). Nevertheless, in the new HBS both the computer-assisted interviewing and the availability of the list of residents from the municipality

Table 1. Main characteristics of the former and new HBS (differences are shown in bold).

	Former HBS		New HBS	
Collected information	All expenditures incurred to directly satisfy household member needs (including self-consumption, imputed rent and gifts).		All expenditures incurred to directly satisfy household member needs (including self-consumption, imputed rent and gifts), as well as the **purchasing month and place**.	
Survey technique	Mixed mode PAPI (1 direct interview) – CADI (self-filled diary)		Mixed mode **CAPI (2 direct interviews)** – CADI (self-filled diary)	
Self-consumption	Amount of produced goods consumed daily and self-assessment of its monetary value, collected by a separate diary delivered only to households declaring self-consumption.		Amount of produced goods consumed daily, **collected by a specific section included in the expenditure diary**. The monetary value is estimated by the **market prices available from consumption prices ISTAT Unit**.	
COICOP classification	COICOP 1993 (producing 265 expenditure items)		COICOP **2013** (producing **482 expenditure items**)	
Reference periods	7 days	110 items (41.5%)	**14 days**	**168 items (34.9%)**
	Last month	90 items (34.0%)	Last month	**54 items (11.2%)**
	Last 3 months	62 items (23.4%)	Last 3 months	**68 items (14.1%)**
	Last 12 months	3 items (1.1%)	Last 12 months	**192 items (39.8%)**
	Total	265 items (100.0%)	Total	**482 items (100.0%)**
Interviewer network	Interviewers selected by municipality statistical offices, trained by regional ISTAT offices.		**Professional interviewers selected by a private company, trained by the central ISTAT office**.	

Source: Grassi and Pannuzi 2015.

register offices allow the substitution rules to be stricter than in the former HBS, which should result in the substitute household being more like the nonresponding household.

For each listed item, in addition to collecting all expenditures incurred by the household members to directly satisfy their needs, the new HBS collects information on: (1) expenditure occurrence; (2) purchasing month; and (3) purchasing location. The first information – about whether or not a given expenditure has been made – makes it possible to distinguish between missing expenditures due to undeclared amounts (to be imputed) and actually null expenditures.

In general, the expenditure reference periods in the new HBS are longer than in the former HBS. Moreover, the percentage of expenditures collected for the previous month has decreased from 34% to 11.2%, and the percentage of expenditures collected for the previous 12 months has increased from 1.1% to 39.8%.

Finally, expenditures for 43 new items, referring to goods and services recently introduced into the market, have been added in the questionnaire. For example, the new items include expenditures for drones or satellite navigators, e-books and e-readers, private security services, or rental of furniture.

2.3. Coverage of Selected Subpopulations

With the aim of evaluating the coverage of selected subpopulations in both surveys, the final and unweighted estimates of the proportions of certain subgroups in both HBSs are compared with the same proportions in the resident population (Table 2). These population subgroups are those involved in the new HBS additional post-stratification procedure, therefore the relative standard errors for the new HBS final estimates are not reported because they are, by definition, equal to 0. These results complement the information already discussed in Lemmi et al. (2019), which proved the better coverage of households by economic condition in the new HBS (compared with the former HBS).

If we consider citizenship, the estimated proportion of foreign people in the former HBS (5.3%) is significantly lower than the proportion in the population (7.8%) (Table 2 panel a, Total population). Moreover, the unweighted estimate (based on the achieved sample) in the new HBS (5.4%) is also closer to the population proportion than the estimates in the former HBS (4.7%). Similar results are observed for the population estimates by employment/professional status. As an example, with reference to managers and white collar employees, compared with the population proportion equal to 14.5%, the unweighted estimated proportions in the former HBS and in the new HBS are 16.9% and 15.9%, respectively, whereas the final estimated proportion in the former HBS is 17.3% (Table 2 – total population estimates).

At the household level, the final estimates show how the households living in the largest municipalities are well covered in both surveys. In terms of unweighted estimates, the new HBS estimates are again closer to the population proportions than the former HBS estimates. The final estimates in the former HBS over-represent the households living in small municipalities and under-represent those living in metropolitan area suburbs and municipalities with more than 50,000 inhabitants (Table 2, panel b, Total households).

These findings confirmed the non-negligible gain of introducing tighter requirements for substitutions and extra post-stratification adjustments to survey control totals in the new survey.

Table 2. Distribution of additional post-stratification variables in the population in the former and new HBS. 2013 (final and unweighted estimate, relative standard errors).

Variables	Population (% distribution)*	Former HBS			New HBS	
		Final estimate			Final estimate	
		% Distribution	Relative standard error (%)	Unweighted estimate % distribution	% Distribution	Unweighted estimate % distribution
a) Total population	**100**	**100**	**–**	**100**	**100**	**100**
By citizenship						
Foreign people	7.8	5.3	4.47	4.7	7.8	5.4
Italian people	92.2	94.7	0.25	95.3	92.2	94.6
By employment/professional condition						
Employees: managers and white collars	14.5	17.3	1.35	16.9	14.5	15.9
Employees: blue collars and similar	13.2	12.7	1.68	12.2	13.2	12.3
Independent: entrepreneurs and freelancers	2.6	2.8	3.65	2.8	2.6	3.4
Independent: self-employed	6.6	4.4	3.03	4.4	6.6	5.8
Looking for a job	5.1	7.5	2.10	7.4	5.1	6.2
Retired from work	20.1	20.4	0.72	22.4	20.1	21.5
In other condition	38.0	34.8	0.60	33.9	38.0	34.8
b) Total households	**100**	**100**	**–**	**100**	**100**	**100**
By municipality typology						
Metropolitan area-centre (municipalities with more than 250,000 inhabitants)	16.8	16.8	1.63	10.9	16.8	11.7
Metropolitan area suburbs and municipalities with more than 50,000 inhabitants	29.3	27.5	3.60	29.7	29.3	29.1
Other municipalities up to 50,000 inhabitants	53.9	55.7	1.74	59.5	53.9	59.3

*The source for the population distribution by citizenship and municipality typology is the total survey on resident population; the source for the population distribution by employment/professional condition is the Labour Force Survey.

2.4. Comparison of Selected Estimates

The survey redesign produced higher consumption expenditure levels together with a smaller range of inequality in the new HBS than in the former HBS. For both surveys, Table 3 shows the estimates of the average consumption expenditure, the deciles of the distribution and the inequality measures (Gini index and income quintile share ratio S80/S20) together with their relative standard errors and 95% confidence intervals calculated on the total expenditure in equivalent terms. The equivalent household consumption expenditure allows comparisons of households with different size or composition. It is obtained using an equivalence scale constituted by a set of coefficients accounting for the economies of scale that can be realised as the household composition changes; the equivalence scale used in this article is known as the Carbonaro scale (Commissione di indagine sulla povertà e sull'emarginazione 1996).

Although the new HBS has a markedly smaller achieved sample size than the former HBS (61% versus 74% of the selected sample), the sampling errors associated with the new HBS estimates are lower than those associated with the former HBS estimates. Moreover, the differences between the estimates obtained by the surveys are statistically significant.

In particular, the difference in the average monthly household consumption expenditure between the two surveys is significantly different from zero and is equal to 3.1%.

The consumption expenditure estimates obtained using the new HBS are higher in value than those obtained using the former HBS along the entire consumption expenditure distribution. The differences are statistically significant for each decile of the equivalent distribution, although they are more marked in the bottom portion of the distribution (for the first two deciles the differences are 13% and 9.8%, respectively). The higher increase in the bottom part of the distribution entails a significantly lower inequality of the consumption expenditure distribution in the new HBS than in the former HBS. The value of the Gini index changes from 0.327 (former HBS) to 0.304 (new HBS), whereas the quintile share ratio – S80/S20 – changes from 5.4 to 4.8 (Table 3).

3. Effect of Single Innovations on Estimates

In this section, we evaluate the effect of each innovation introduced in the new HBS on estimates of levels and inequality. We are interested in isolating the effects of single changes because they might cancel each other out when the global effect is considered. In certain cases, our findings simply confirm what is already contained in the specific literature, whereas in other cases, the evidence might represent new elements to be considered when HBSs or households sample surveys in general must be redesigned. As advised by the relevant literature (Van den Brakel et al. 2017; Gazzelloni 2006; Zbikowski and Lubich 2006; Polivka and Miller 1998), the HBS redesign process was preceded by a series of experimental surveys (Grassi and Pannuzi 2015). Furthermore, the former and new HBSs were conducted in parallel during the last two quarters of 2012 and all quarters of 2013. To assess the effects of innovations on the estimates, we have conducted a detailed *ad hoc* simulation.

Table 3. Average, decile values, Gini and S80/S20 inequality indicators for the equivalent consumption expenditure in the former and new HBS. 2013 (values and standard errors).

	Former HBS				New HBS				
			95% confidence intervals				95% confidence intervals		
Indicator	Value (a)	Relative standard error (%)	Lower	Upper	Value (b)	Relative standard error (%)	Lower	Upper	% Difference [(b-a)/a]*100
Average	2,366	0.65	2,336	2,396	2,440	0.58	2,412	2,468	3.1
1 decile	906	1.09	888	924	1,024	1.02	1,002	1,046	13.0
2 decile	1,201	0.76	1,177	1,224	1,318	0.99	1,299	1,338	9.8
3 decile	1,498	0.71	1,474	1,522	1,579	0.83	1,557	1,601	5.4
4 decile	1,763	0.62	1,743	1,782	1,825	0.56	1,803	1,847	3.5
5 decile	1,997	0.75	1,974	2,021	2,096	0.61	2,065	2,127	4.9
6 decile	2,284	0.68	2,258	2,310	2,424	0.58	2,392	2,457	6.1
7 decile	2,613	0.69	2,585	2,641	2,809	0.54	2,771	2,847	7.5
8 decile	3,067	0.75	3,008	3,125	3,338	0.97	3,289	3,388	8.9
9 decile	4,086	1.01	4,028	4,144	4,233	0.72	4,149	4,317	3.6
Gini	0.327	0.95	0.321	0.333	0.304	0.79	0.299	0.309	–
S80/S20	5.4	0.02	5.4	5.4	4.8	0.01	4.8	4.8	–

Note: The standard errors of the deciles and inequality measures were obtained by the generalized linearization method proposed by Osier (2009) for the Eurostat Laeken complex indicators, which relies on the concept of the influence function.

Starting from the data of the new HBS, we performed the following steps:

1. we considered for each item only the expenditures made in the portion of the reference period overlapping the former HBS reference period,
2. we disregarded all the imputed expenditures due to undeclared amount,
3. we aggregated the expenditures referring to items that have been split in the new HBS (3.1) and we ignored the expenditures for new items that were not included in the former HBS (3.2), and
4. we recalculated the weights without the post-stratification adjustments to survey control totals introduced in the new HBS.

We first comment on the findings with reference to the estimate of the average consumption expenditure (Table 4). For step (1), because the reference periods in the new HBS are generally larger than in the former HBS, we recalculated the average consumption expenditure on the new HBS data for each item obtained by considering only the expenditures afforded during the reference period used in the former HBS for the same item. For example, in the former HBS, the expenditures for domestic services were collected with reference to the last month, whereas in the new HBS they are collected with reference to the last three months. For our purposes, we recalculated the average monthly expenditures on new HBS data by considering only the expenditures for domestic services made in the last month. This recalculation was possible due to the information

Table 4. Simulation of the redesign effects on average consumption expenditure. 2013 (values in euros and percentage variation).

Data source	Average equivalent consumption expenditure	% Variation to the previous value
Former HBS	2,366	
New HBS	2,440	3.1
1. New HBS with the former HBS reference periods	2,469	1.2
2. New HBS with the former HBS reference periods and without imputed expenditures	2,421	− 1.9
3.1 New HBS with the former HBS reference periods, without imputed expenditures and without the splitting effect	2,361	− 2.5
3.2 New HBS with the former HBS reference periods, without imputed expenditures, without the splitting effect and without the new items	2,244	− 5.0
4. New HBS with the former HBS reference periods, without imputed expenditures, without splitting effect, without the new items and without new post-stratification constraints *(new HBS "treated to simulate" the former HBS methodology)*	2,341	4.3

on the purchasing month collected in the new HBS. By applying this process for all items, the average monthly consumption expenditure increases from EUR 2,440 to EUR 2,469, which corresponds to a change of +1.2%. In other words, using longer reference periods results in lower average consumption expenditure estimates. This result could be due to the "memory recall error", namely, that households tend to forget expenditures as the length of the reference period increases. Nevertheless, in the new HBS, this effect has been kept under control by adopting specific strategies, as indicated in the literature (Mulry et al. 2016; Mathiowetz et al. 2002; Neter and Waksberg 1964). In particular, in addition to precise anchoring of the time frame and precise definitions, we introduced an easier-to-answer set of questions related to the event of interest ("warm-up" questions), referring to whether the expenditure was made, in which month it was made, and the purchasing place. From the answers to the "warm up" questions, we have also been able to impute the missing expenditures. The expenditures, similarly to other quantitative variables, are treated by IVEware software (developed by the Survey Research Center, Institute for Social Research, University of Michigan), which makes single or multiple imputation by model using the sequential regressions method (Raghunathan et al. 2001) and Banff software (developed by Statistics Canada), based on Fellegi-Holt methodology through the Nearest-Neighbor Donor method (see Grassi and Pannuzi 2015).

For our *ad hoc* simulation, with reference to step 2, we recalculated the average consumption expenditure on the data from the previous step without considering the imputed expenditures, and the value changes from EUR 2,469 to EUR 2,421 (−1.9%). This result suggests that using expenditure-related (easy to answer) questions (as in the new HBS) allows us to distinguish between a missing expenditure and actually null expenditure, and in this manner, imputation can be used to recover missing values and complete the estimated consumption expenditure (Gonzalez and Eltinge 2010).

Step 3 includes the effect of splitting items (3.1) and the effect due to the introduction of new items (3.2). For the former point, because the COICOP version adopted in the new HBS is more detailed than that used in the former HBS, for the data from step 2, we aggregated items and summed all of the related expenditures to reproduce the same classification used in the former HBS. To give an example, the aggregate "Bread, breadsticks and crackers", which corresponds to a single item in the former HBS, is obtained by summing two items in the new HBS: "Bread" and "Breadsticks and crackers". By comparing the expenditure referred to these aggregates, we can estimate the effect of the item splitting. The consumption expenditure changes from EUR 2,421 to EUR 2,361. This result is consistent with what is already known in the literature (Crossley and Winter 2015), that a greater disaggregation of items induces higher estimates of the total expenditure, presumably because it helps households to remember expenses that could otherwise be forgotten (Cifaldi and Neri 2013).

The periodic updating of the COICOP classification also has the aim of following the change in goods and services available in the market by eliminating those that are no longer sold and introducing novelties instead. To assess the effect of the introduction of new items (step 3.2), we exclude the expenditures referred to new items from the data coming from step 3.1. This simulation obviously produces a reduction of the monthly expenditure, specifically from EUR 2,361 to EUR 2,244 (−5.0%).

Finally, (step 4), because additional post-stratification adjustments have been added in the weighting system in the new HBS, we recalculated the weights without the new post-stratification adjustments (in a similar manner as in the former HBS). The consumption expenditure calculated using data from step 3.2 and the weights without the post-stratification adjustments increases from EUR 2,244 to EUR 2,341 (+4.3%). This effect is strictly linked to the nature of the variables used in post-stratification. In our case, the new adjustments give more weight to subgroups of the population with low levels of consumption expenditure (foreign people in particular). As already known in the literature, introducing adjustments in the final weight calculation allows us to improve the accuracy of the estimates, and the more the auxiliary variables considered are associated with the variables under investigation, the more the distortion of the estimates is reduced (Lavallé and Beaumont 2015).

The last estimate (point 4 of Table 4) represents the final result of the simulation, and specifically, it is the average consumption expenditure from the *new HBS "treated to simulate" the former HBS* methodology according to the above-described sequence of steps. Considering the associated relative standard errors (equal to 0.76% for the *new HBS "treated to simulate" the former HBS*), it is statistically identical to the estimated average based on the former HBS data.

The simulation of the redesign effect on inequality measures along the abovementioned steps 1 to 4 (Table 5) shows that the introduction of new items is the only innovation that produces an increase in inequality, whereas the other innovations tend to reduce the degree of inequality. This result appears to be linked to the fact that the new items refer to expenditures for goods and services that are more likely afforded by households/people with high levels of consumption. However, in terms of inequality, the estimates calculated

Table 5. Simulation of the redesign effects on inequality indices. 2013.

Data source	Inequality index (S80/S20)	Gini index
Former HBS	5.4	0.327
New HBS	4.8	0.304
1. New HBS with the former HBS reference periods	5.2	0.324
2. New HBS with the former HBS reference periods and without imputed expenditures	5.4	0.330
3.1 New HBS with the former HBS reference periods, without imputed expenditures and without the splitting effect	5.4	0.333
3.2 New HBS with the former HBS reference periods, without imputed expenditures, without the splitting effect and without the new items	5.2	0.329
4. New HBS with the former HBS reference periods, without imputed expenditures, without splitting effect, without the new items and without new post-stratification constraints *(new HBS "treated to simulate" the former HBS methodology)*	5.5	0.334

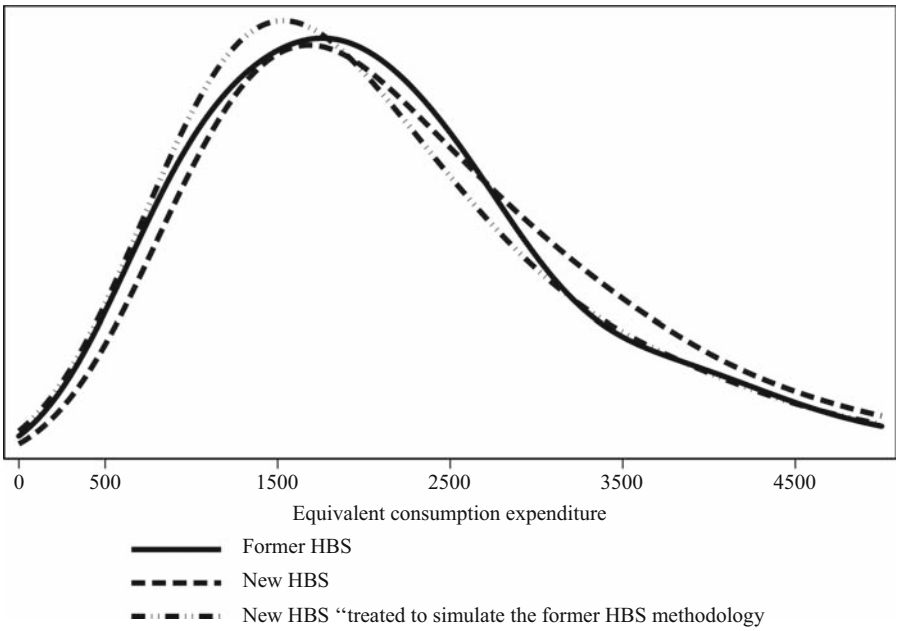


Fig. 1. Kernel density estimation of the equivalent consumption expenditure distribution in the former HBS, new HBS and new HBS "treated to simulate" the former HBS methodology. 2013.

on the *new HBS "treated to simulate" the former HBS* data are also closer to those resulting from the former HBS data than to those obtained on new HBS data.

The "simulation" in Table 5 does not presume to reproduce the former HBS data via the new HBS data. Indeed, several sources of difference have not been included in the simulation; for example, the effect of the survey technique change from PAPI to CAPI or of the interviewer network change from public to private/professional. Nevertheless, even if the distributions are different (this is confirmed by the Kolmogorov-Smirnov and Kuiper tests, which, however, tend to be rather conservative with large samples), it is undeniable that the equivalent consumption expenditure distribution on data from the former HBS is more similar to that obtained on data from the *new HBS "treated to simulate" the former HBS* than that obtained on data from the new HBS, especially if considering the distribution tails (Figure 1).

4. Association Between Household Characteristics and Consumption Expenditure Level

This section investigates whether the new HBS redesign has modified the consumption expenditure level estimates in a different way for different household subgroups.

Table 6 (see Appendix, Section 6) reports the household percentage composition, the average equivalent consumption expenditure and its relative standard error (including the 95% confidence intervals) for selected subgroups of households from both surveys. The last column contains the estimated percentage difference with assessment of its significance via the two-sample t-test. The average consumption estimate in the new HBS is significantly higher than in the former HBS for (1) households living in either the Centre

or South and Islands, (2) municipalities in metropolitan area (both centre or suburbs) and with at least 50,000 inhabitants, (3) households with at least two members, (4) households with no minor children, (5) households with elderly members, (6) households headed by a retired person or by an employee in non-manual jobs, or (7) households without self-consumption. The largest differences (slightly less than 10%) are observed for households with at least four members and for those living in metropolitan area centres. No subgroup of households exhibits a significantly higher average consumption expenditure in the former HBS than in the new HBS.

The consumption expenditure levels are linked with both the expenditure amount and expenditure frequency, which in turn are strongly associated with household size. The enlargement of the reference periods, as well as the imputation of expenditures and the introduction of new items in the new HBS has produced an increase in the reported number of expenditure events. Figure 2 shows that this increase is more pronounced for larger households (28% for households with four or more members versus 21% among single persons).

As a consequence, in the new HBS compared with the old HBS, we observe an increase in the share of small households and a reduction of the share of large households in the bottom fifth of the distribution. Moreover, this effect is more evident among households in the South and Islands and households in the largest municipalities (Figure 3).

This evidence appears to be justified by the changing purchasing behaviours. In recent decades, households have reduced the time spent shopping and have made less frequent shopping trips. The opportunities to purchase larger quantities of goods at lower unit prices or to benefit from discounted prices encourage households to buy goods in bulk and to stock up in ways that were not feasible in the past (Censis 2012).

It is clear that this behaviour more frequently belongs to large households (those able to buy and consume larger quantities of goods), households in metropolitan areas (where the mobility costs are higher than in smaller municipalities) or in the South and Islands (where households more often have stricter budget constraints that might impose saving strategies) (Bank of Italy 2018).

Fig. 2. Average number of expenditure events by household size in the former and new HBS. 2013.

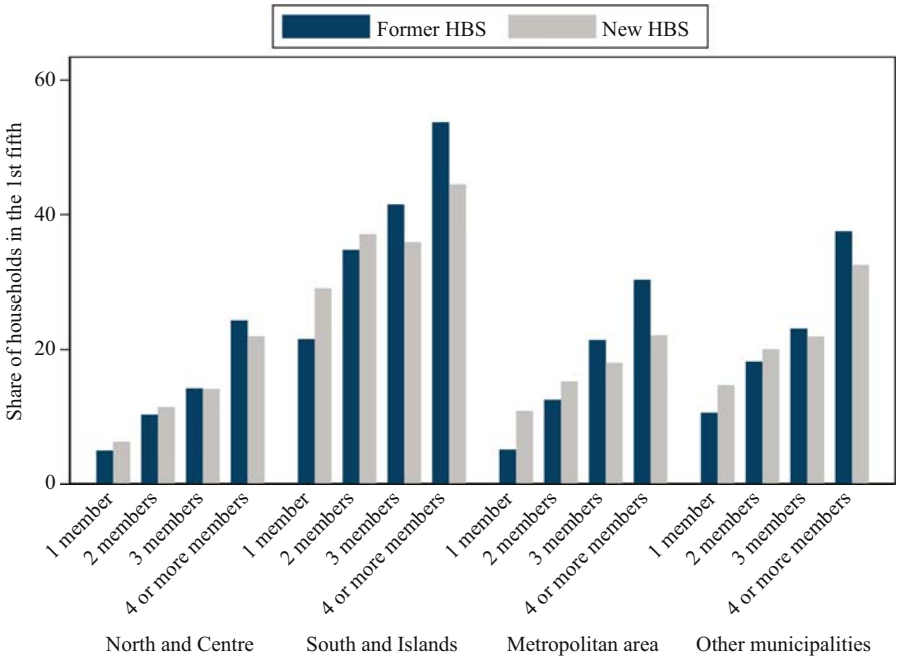


Fig. 3. Share of households in the first fifth of the equivalent consumption expenditure distribution by household size and geographical area and by household size and municipality typology in the former and new HBS. 2013.

Moreover, the self-consumption collection mode in the new HBS also produces a higher impact on estimates for households in the South and on Islands than elsewhere. Comparing the former and the new HBS, the share of households with self-consumption in both surveys remains nearly the same in the North (around 6.2%), doubles in the Centre (reaching 8.7%) and nearly triples in the South and on the Islands, where the share exceeds 10.3% in the new HBS.

Finally, to supply evidence of the main determinants of consumption expenditure levels in both surveys, regression models have been estimated.

The dependent variable is defined by taking logarithmic transformation of the consumption expenditures. The covariates are the following household characteristics, all entered as binary (0/1) variables: geographical area, household size, presence of elderly or minor members, municipality typology, occupational/professional status of the household head, and presence of self-consumption. Using the logarithmic transformation of the dependent variable, the estimated coefficient associated with a given category of the binary predictor approximates the relative gap in the average consumption expenditure with respect to the reference category when the other covariates are held constant.

The linear regression model estimates (Appendix, Table 7) show how the consumption levels are higher in both surveys for households living in the North, in a metropolitan area centre, with five or more members, without minor or elderly members, or headed by an employer or professional. However, the lowest levels are registered for households living in the South, in small municipalities, with only one member or headed by someone looking for a job. In the new HBS, households with self-consumption exhibit a significantly higher level of consumption compared with households with the same characteristics not

reporting self-consumption. In the former HBS, this effect is not significant. This result is linked to the different way of collecting self-consumption in the new HBS, which increased the share of households with self-consumption. If we analyse the results of the quantile regression (Table 7 in Appendix reports the estimates for the 25th, 50th and 75th percentiles), we can confirm that along the entire distribution, the consumption levels are higher for households living in the North, in the largest municipalities, without minor children or elderly members, or headed by an employer or professional. Generally, the highest levels of consumption for large households are also confirmed for the new HBS and especially in the bottom portion of the distribution, even if the difference between households with four and households with five or more members is not statistically significant. As previously noted for the linear model, in the new HBS, households with self-consumption have a higher level of expenditure than the other households.

5. Conclusions

Controlling all of the possible sources of errors is a crucial aspect for survey design or renewal. In particular, in this article, we assessed the effect of the innovations introduced in the new Italian HBS, and we found that the new survey produced estimates closer to those obtained by external sources with smaller sample errors in comparison with the former HBS.

Specifically, despite a smaller achieved sample size, the new survey design has succeeded in reducing the variance of the average, the vast majority of the deciles and the inequality measures of the equivalent consumption expenditure. The decrease in variance is especially strong for estimates of the average consumption expenditure by municipality typology, which has become a new domain for stratification of the units at the first stage in the new HBS. We also showed a non-negligible gain in terms of coverage error reduction due to extra post-stratification adjustments to citizenship and professional condition of the population and to the typology of household municipality of residence, together with tighter requirements for substitutions to replace nonresponding households.

As the result of all of the changes introduced in the transition from the former survey to the new survey, the overall effect is that the new HBS exhibits higher levels of consumption expenditure, especially in the bottom portion of the distribution, and a lower inequality degree. These differences are all statistically significant.

The main novelty of the article lies in investigation of the effects of changes in specific aspects of the survey, which are usually less addressed in the literature. Indeed, the effect of survey redesign has been deeply studied in the specific literature, primarily with reference to such aspects as survey technique and interviewer effect, to name a few. Conducting the surveys in parallel enabled separate quantification of the impact due to the enlargement of the reference period for the expenditures, the imputation of zero expenditures, the splitting of expenditure items, the introduction of new items and the addition of post-stratification adjustments. We found that changes in these aspects have a non-negligible effect on the final estimates.

In particular, the results show how imputing zero expenditures, splitting items of consumption and introducing new items combine to produce an increase in the estimated consumption expenditure levels, as expected. Moreover, the increase due to all of these

changes more than compensates for the decrease in the consumption expenditure due to the introduction of post-stratification adjustments and the extension of reference periods.

With reference to the effects of the redesign on the inequality measures, we found that the introduction of new expenditure items is the only inequality-enhancing change, whereas every other change tends to reduce the inequality degree. The sign of this effect depends on which subgroups of population are primarily involved in the purchase of the new goods or services. In the Italian HBS, the expenditures for new items occur mainly in households with a high level of consumption expenditure, which explains the increase in inequality.

The effect of the introduction of new post-stratification depends on which subgroups of the population are better represented after weighting. In the Italian HBS, the weight system increases the share of population with a low level of consumption, producing a decrease in both level and inequality of consumption expenditure.

The enlargement of the reference periods allows both a better approximation of the household consumption expenditure and an increase in the probability of capturing the expenditure event, thus reducing the proportion of households with no expenditures and, therefore, the variability of the estimates. Nevertheless, larger periods might give rise to strong memory recall errors; therefore, it is crucial to use strategies to control or reduce them.

Although the difference between the population average consumption expenditure in the former and new HBS does not appear to be remarkable, even if statistically significant, it is particularly marked for large households, those living in the South or on Islands, or in metropolitan areas, with a relevant impact on variability and inequality. Moreover, the new HBS is also able to capture the effect of self-consumption, given the better instrument used to collect the information, which increased the number of households with self-consumption.

6. Appendix

Table 6. Household percentage composition, average equivalent consumption expenditure in the former and new HBS by household characteristics. Year 2013 (values in euro, relative standard errors and percentage difference).

	Former HBS					New HBS					
		Equivalent consumption expenditure					Equivalent consumption expenditure				
Household characteristics	% HH composition	Average (a)	Relative standard error (%)	95% Confidence intervals Lower	95% Confidence intervals Upper	% HH composition	Average (b)	Relative standard error (%)	95% Confidence intervals Lower	95% Confidence intervals Upper	% Difference [(b-a)/a]*100
Geographical area											
North	48.1	2,764	0.89	2,716	2,812	47.9	2,789	0.80	2,745	2,833	0.9
Centre	20.4	2,486	1.53	2,411	2,561	20.3	2,638	1.47	2,562	2,714	6.1*
South and Islands	31.5	1,679	1.16	1,641	1,717	31.9	1,789	1.00	1,754	1,824	6.5*
Municipality typology											
Metropolitan area-centre	16.8	2,685	1.50	2,606	2,764	16.8	2,945	1.45	2,861	3,029	9.7*
Metropolitan area suburbs and municipalities with more than 50,000 inhabitants	27.5	2,403	3.65	2,231	2,575	29.3	2,507	1.06	2,455	2,559	4.3*
Other municipalities up to 50,000 inhabitants	55.7	2,251	2.00	2,163	2,339	53.9	2,246	0.78	2,212	2,280	−0.2
Household size											
HH with 1 member	32.7	2,981	1.42	2,898	3,064	31.6	2,951	1.33	2,874	3,028	−1.0
HH with 2 members	26.5	2,389	1.63	2,313	2,465	29.2	2,485	1.70	2,402	2,568	4.0*
HH with 3 members	18.9	2,063	2.20	1,974	2,152	17.4	2,192	2.06	2,103	2,281	6.2*
HH with 4 members	16.9	1,745	2.11	1,673	1,817	16.7	1,917	1.98	1,843	1,991	9.8*
HH with 5 or more members	5.0	1,451	4.07	1,335	1,567	5.2	1,579	3.63	1,467	1,691	8.8*
Household with minor children											
HH without minor children	75.1	2,537	0.76	2,499	2,575	74.9	2,612	0.74	2,574	2,650	2.9*
HH with 1 minor child	13.5	2,005	2.53	1,906	2,104	13.0	2,076	2.65	1,968	2,184	3.5
HH with 2 or more minor children	11.4	1,662	2.16	1,592	1,732	12.1	1,766	2.08	1,694	1,838	6.3
Household with elderly											
HH without elderly	63.0	2,377	0.82	2,339	2,415	63.0	2,422	0.94	2,377	2,467	1.9
HH with 1 elderly	24.9	2,450	1.99	2,354	2,546	24.2	2,564	1.76	2,476	2,652	4.7*
HH with 2 or more elderly	12.1	2,130	2.23	2,037	2,223	12.7	2,292	2.06	2,199	2,385	7.6*

Table 6. Continued.

Household characteristics	Former HBS					New HBS					
		Equivalent consumption expenditure					Equivalent consumption expenditure				
	% HH composition	Average (a)	Relative standard error (%)	95% Confidence intervals		% HH composition	Average (b)	Relative standard error (%)	95% Confidence intervals		% Difference [(b-a)/a]*100
				Lower	Upper				Lower	Upper	
Reference person professional/occupational condition											
Employees: managers and white collars	21.7	2,721	2.11	2,608	2,834	18.9	2,867	1.61	2,777	2,957	5.4*
Employees: blue collars and similar	18.8	1,962	2.38	1,870	2,054	18.4	2,001	1.69	1,935	2,067	2.0
Independent: entrepreneurs and freelancers	4.5	3,105	4.96	2,803	3,407	4.4	3,394	3.56	3,157	3,631	9.3
Independent: self-employed	6.8	2,374	4.89	2,146	2,602	9.9	2,379	2.56	2,260	2,498	0.2
Looking for a job	4.4	1,748	5.59	1,556	1,940	3.2	1,677	5.37	1,500	1,854	−4.1
Retired from work	35.8	2,405	1.20	2,348	2,462	35.4	2,505	1.37	2,438	2,572	4.2*
In other condition	8.0	2,090	3.95	1,928	2,252	9.8	2,095	3.65	1,945	2,245	0.2
Self-consumption											
With self-consumption	5.0	2,339	5.73	2,076	2,602	8.0	2,390	4.11	2,197	2,583	2.2
Without self-consumption	95.0	2,367	0.75	2,332	2,402	92.0	2,444	0.66	2,412	2,476	3.3*

*: significant at 5% level.

Table 7. Results of linear and quantile regression models for household consumption expenditure (logarithmic transformation) in the former and new HBS, Year 2013.

	Former HBS				New HBS			
	Linear regression estimates	Quantile regression estimates (at selected quantiles)			Linear regression estimates	Quantile regression estimates (at selected quantiles)		
Covariates		25th	50th	75th		25th	50th	75th
Intercept	7.125***	6.69***	7.16***	7.47***	7.18***	6.83***	7.2***	7.56***
Geographical area (ref = South and Islands)								
North	0.421***	0.45***	0.41***	0.39***	0.398***	0.41***	0.4***	0.39***
Centre	0.326***	0.39***	0.33***	0.28***	0.298***	0.31***	0.29***	0.27***
Municipality typology (ref: Other municipalities up to 50,000 inhabitants)								
Metropolitan area-centre	0.116***	0.12***	0.11***	0.13***	0.183***	0.21***	0.21***	0.18***
Metropolitan area suburbs and municipalities with more than 50,000 inhabitants	0.071***	0.07***	0.06***	0.08***	0.094***	0.09***	0.08***	0.09***
Household size (ref = HH of 5 or more members)								
HH with 1 member	−0.6***	−0.6***	−0.63***	−0.61***	−0.742***	−0.78***	−0.8***	−0.79***
HH with 2 members	−0.281***	−0.25***	−0.3***	−0.32***	−0.388***	−0.4***	−0.45***	−0.42***
HH with 3 members	−0.132***	−0.11***	−0.15***	−0.14***	−0.185***	−0.18***	−0.22***	−0.24***
HH with 4 members	−0.039**	−0.01	−0.04*	−0.05*	−0.048***	−0.04	−0.1***	−0.09***
Household with minor children (ref = HH with 2 or more children)								
HH without minor children	0.13***	0.13***	0.13***	0.14***	0.154***	0.15***	0.15***	0.19***
HH with 1 minor child	0.063***	0.06***	0.06***	0.08***	0.072***	0.07**	0.06**	0.1***
Household with elderly (ref = HH with 2 or more elderly)								
HH without elderly	0.113***	0.13***	0.12***	0.13***	0.102***	0.12***	0.15***	0.12***
HH with 1 elderly	0.008	0.03*	0	0.02	0.034**	0.03	0.05**	0.04**

Table 7. Continued.

Covariates	Former HBS				New HBS			
	Linear regression estimates	Quantile regression estimates (at selected quantiles)			Linear regression estimates	Quantile regression estimates (at selected quantiles)		
		25th	50th	75th		25th	50th	75th
Reference person professional/occupational condition (ref = job seeker)								
Employees: managers and white collars	0.445***	0.53***	0.39***	0.39***	0.527***	0.61***	0.51***	0.43***
Employees: blue collars and similar	0.169***	0.24***	0.14***	0.13***	0.211***	0.25***	0.22***	0.16***
Independent: entrepreneurs and freelancers	0.558***	0.6***	0.52***	0.53***	0.659***	0.69***	0.63***	0.57***
Independent: self-employed	0.315***	0.39***	0.27***	0.27***	0.371***	0.38***	0.38***	0.32***
Retired from work	0.268***	0.34***	0.26***	0.25***	0.334***	0.36***	0.35***	0.29***
In other condition	0.141***	0.22***	0.14***	0.1***	0.185***	0.22***	0.18***	0.17***
Self-consumption (ref = No)								
Yes	0.011	−0.01	−0.02	0.04*	0.133***	0.15***	0.14***	0.11***
Adjusted R-squared	*0.304*				*0.358*			

*significant at 10% level.
**significant at 5% level.
***significant at 1% level.
Note: Weighted regressions have been performed by the SAS procedures REG and QUANTREG. For the linear regression, we have verified that controlling for the stratification in the survey design through the SAS procedure SURVEYREG does not change our conclusions: specifically, the coefficients that are not significant in the weighted regression still remain not significant.

7. References

Andreski, P., G. Li, M.Z. Samancioglu, and R. Schoeni. 2014. "Estimates of Annual Consumption Expenditures and Its Major Components in the PSID in Comparison to the CE." *American Economic Review* 104(5): 132–135. DOI: https://doi.org/10.1257/aer.104.5.132.

Angrisani, M., A. Kapteyn, and S. Schuh. 2015. "Measuring Household Spending and Payment Habits: The Role of "Typical" and "Specific" Time Frames in Survey Questions." In *Improving the Measurement of Consumer Expenditures*, edited by C.D. Carroll, T.F. Crossley, and J. Sabelhaus. NBER Book Series Studies in Income and Wealth, 74: 414–440. Chicago: University Press. DOI: https://doi.org/10.7208/chicago/9780226194714.003.0016.

Bank of Italy. 2018. *Annual report for 2017, 124th Financial Year*, Rome. Available at: https://www.bancaditalia.it/pubblicazioni/relazione-annuale/2017/en_rel_2017.pdf?language_id=1 (accessed March 2019).

Barrett, G., P. Levell, and K. Milligan. 2015. "A comparison of micro and macro expenditure measures across countries using differing survey methods." In *Improving the Measurement of Consumer Expenditures*, edited by C.D. Carroll, T.F. Crossley, and J. Sabelhaus. NBER Book Series Studies in Income and Wealth, 74: 263–286. Chicago: University of Chicago Press. DOI: https://doi.org/10.7208/chicago/9780226194714.003.0010.

Bee, A., B.D. Meyer, and J.X. Sullivan. 2015. "The validity of consumption data: Are the consumer expenditure interview and diary surveys informative?". In *Improving the Measurement of Consumer Expenditures*, edited by C.D. Carroll, T.F. Crossley, and J. Sabelhaus. NBER Book Series Studies in Income and Wealth, 74: 204–240. Chicago: University Press. DOI: https://doi.org/10.7208/chicago/9780226194714.003.0008.

Censis. 2012. *Crisi: Vivere insieme, vivere meglio.* Ricerca Censis-Coldiretti. Centro Studi Investimenti Sociali. Available at: https://www.coldiretti.it/archivio/il-rapporto-coldiretticensis-crisi-vivere-insieme-vivere-meglio (accessed April 2018).

Cifaldi, G. and A. Neri. 2013. "Asking income and consumption questions in the same survey: what are the risks?". *Bank of Italy Economic Working Papers:* 908. Available at: https://www.bancaditalia.it/pubblicazioni/temidiscussione/2013/2013-0908/en_tema_908.pdf (accessed March 2018).

Commissione di indagine sulla povertà e sull'emarginazione. 1996. "Le misure della povertà in Italia: scale di equivalenza e aspetti demografici." Presidenza del Consiglio dei Ministri. Roma. Available at: http://sitiarcheologici.lavoro.gov.it/Documents/Resources/Lavoro/CIES/Scale_equivalenza_1996.pdf (accessed October 2018).

Crossley, T.F. and J.K. Winter. 2015. "Asking Households About Expenditures: What Have We Learned?". In *Improving the Measurement of Consumer Expenditures*, edited by C.D. Carroll, T.F. Crossley, and J. Sabelhaus. NBER Book Series Studies in Income and Wealth, 74: 23–50. Chicago: University Press. DOI: https://doi.org/10.7208/chicago/9780226194714.003.0002.

Eurostat. 2003. *Household Budget Surveys in the EU. Methodology and Recommendations for Harmonisation–2003. Methods and Nomenclatures*. Luxembourg: Eurostat.

Available at: http://ec.europa.eu/eurostat/ramon/statmanuals/files/KS-BF-03-003-__-N-EN.pdf (accessed February 2018).

Eurostat. 2013. *COICOP Five-Digit – Structure and Explanatory Notes, Unit B5*. Luxembourg. Available at: https://www.dst.dk/ext/4197663288/0/pris/COICOP−pdf (accessed February 2018).

Eurostat. 2015. *Household Budget Survey, 2010 Wave. EU Quality report*. Luxembourg. Available at: http://ec.europa.eu/eurostat/documents/54431/1966394/2015-04-01_QualityReport2010.pdf/418a037a-bfbc-486e-9ff7-4b140b543f39 (accessed February 2018).

Eurostat. 2017. "Transmission, processing and publication of HBS 2015 data." Paper presented at the Working group on Income and Living Conditions, Household Budget Survey, 28 September 2017. Luxembourg. Available at: https://slideplayer.com/slide/15095624/#.Xp3JyPTpZQ0.gmail (accessed February 2018).

Freguja, C. and C. Romano, eds. 2014. "La modernizzazione delle tecniche di rilevazione nelle indagini socio-economiche sulle famiglie". *Metodi-Letture statistiche Istat Series*. ISBN 978-88-458-1806-6. Available at: https://www.istat.it/it/archivio/145721 (accessed March 2018).

Gazzelloni, S., ed. 2006. "La rilevazione sulle forze di lavoro: contenuti, metodologie, organizzazione". *Metodi e Norme Istat Series* 32. ISBN 88-458-1357-6. Available at: https://www.istat.it/it/files/2014/06/met_norme_06_32_-rilevazione_forze_lavoro.pdf (accessed February 2018).

Gonzalez, M.J. and J.L. Eltinge. 2010. Sensitivity of Inference under Imputation: An Empirical Study. Available at: https://www.researchgate.net/publication/265893889_Sensitivity_of_Inference_under_Imputation_An_Empirical_Study (accessed April 2018).

Grassi, D. and N. Pannuzi, eds. 2015. "La nuova indagine sulle spese per consumi in Italia". *Metodi-Letture statistiche Istat Series*. ISBN 978-88-458-1856-1. Available at: https://www.istat.it/it/archivio/182165 (accessed February 2018).

Istat. 2015. *La spesa per consumi delle famiglie. Anno 2014*. 8 July 2015. Available at: https://www.istat.it/it/archivio/164313 (accessed February 2018).

Lavallée, P. and J.-F. Beaumont. 2015. "Why We Should Put Some Weight on Weights". *Survey Insights: Methods from the Field, Weighting: Practical Issues and 'How to' Approach, Invited article*. Available at: https://surveyinsights.org/?p=6255 (accessed March 2019).

Lemmi, A., D. Grassi, A. Masi, N. Pannuzi, and A. Regoli. 2019. "Methodological Choices and Data Quality Issues for Official Poverty Measures: Evidences from Italy". *Social Indicators Research* 141(1): 299−330. DOI: https://doi.org/10.1007/s11205-018-1841-3.

Mathiowetz, N.A., C. Brown, and J. Bound. 2002. "Measurement Error in Surveys of the Low-Income Population". In *Studies of Welfare Populations: Data Collection and Research Issues,* edited by M. Ver Ploeg, R.A. Moffit, and C.F. Citro: 157−194. DC: National Research Council. DOI: https://doi.org/10.17226/10206.

Mulry, M.H, E.M. Nichols, and J.H. Childs. 2016. "A Case Study of Error in Survey Reports of Move Month Using the U.S. Postal Service Change of Address Records" . *Survey Methods: Insights from the Field.* DOI: https://doi.org/10.13094/SMIF-2016-00004.

Neter, J. and J. Waksberg. 1964. "A Study of Response Errors in Expenditures Data from Household Interviews". *Journal of the American Statistical Association* 59(305): 18−55. DOI: https://doi.org/10.1080/01621459.1964.10480699.

Ngandu, N.K., S. Manda, D. Besada, S. Rohde, N.P. Oliphant, and T. Doherty. 2016. "Does adjusting for recall in trend analysis affect coverage estimates for maternal and child health indicators? An analysis of DHS and MICS survey data". *Glob Health Action* 9. DOI: https://doi.org/10.3402/gha.v9.32408.

OECD. 2013. *OECD Framework for Statistics on the Distribution of Household Income, Consumption and Wealth*. OECD Publishing. Paris. DOI: https://doi.org/10.1787/9789264194830-en.

Osier, G. 2009. "Variance estimation for complex indicators of poverty and inequality using linearization techniques". *Survey Research Methods* 3(3): 167–195. DOI: https://doi.org/10.18148/srm/2009.v3i3.369.

Polivka, A.E. and S.M. Miller. 1998. "The CPS after the Redesign: Refocusing the Economic Lens Chapter". In *Labor Statistics Measurement Issues*, edited by John Haltiwanger, Marilyn E. Manser, and Robert Topel, 249–289. Chicago: University of Chicago Press.

Raghunathan, T.E., J.M. Lepkowski, J. van Hoewyk, and P. Solenberger. 2001. "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models". *Survey Methodology* 27(1): 85–95. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2001001/article/5857-eng.pdf?st=HUdPxONq (accessed June 2019).

Smith, L.C., O. Dupriez, and N. Troubat. 2014. "Assessment of the Reliability and Relevance of the Food Data Collected in National Household Consumption and Expenditure Surveys". *International Household Survey Network Working Paper No. 008*. Available at: http://www.ihsn.org/sites/default/files/resources/IHSN_WP008_EN.pdf (accessed April 2018).

UN Statistical Commission. 2018. "Revised Classification of Individual Consumption According to Purpose (COICOP 2018) – Introductory guidelines." Prepared by the Technical Subgroup for the Revision of COICOP (TSG-COICOP), Statistical Commission Background document, Forty-ninth session 6–9 March 2018. Available at: https://unstats.un.org/unsd/statcom/49th-session/documents/BG-Item3l-Classification-E.pdf (accessed February 2018).

Van den Brakel, J., X-M. Zhang, and S.M. Tam. 2017. "Measuring discontinuities due to survey process redesigns". *CBS discussion paper* 2017/13. Statistics Netherlands. Available at: https://www.cbs.nl/en-gb/background/2017/30/measuring-discontinuities-due-to-survey-process-redesigns (accessed February 2018).

Zbikowski, A. and A. Lubich, eds. 2006. "Design and Methodology Current Population Survey U.S. Census Bureau". *Current population Survey Design and Methodology Technical Paper 66*. Available at: https://www.census.gov/prod/2006pubs/tp-66.pdf (accessed February 2018).

Zezza, A., C. Carletto, J.L. Fiedler, P. Gennari, and D. Jolliffe. 2017. Food counts. Measuring food consumption and expenditures in household consumption and expenditure surveys (HCES). Introduction to the special issue. *Food Policy* 72: 1–6. DOI: https://doi.org/10.1016/j.foodpol.2017.08.007.

Received August 2018
Revised October 2019
Accepted January 2020

sciendo

Journal of Official Statistics, Vol. 36, No. 2, 2020, pp. 435–458, http://dx.doi.org/10.2478/JOS-2020-0022

On Accuracy Estimation Using Parametric Bootstrap in small Area Prediction Problems

Tomasz Żądło[1]

We consider longitudinal data and the problem of prediction of subpopulation (domain) characteristics that can be written as a linear combination of the variable of interest, including cases of small or zero sample sizes in the domain and time period of interest. We consider the empirical version of the predictor proposed by Royall (1976) showing that it is a generalization of the empirical version of the predictor presented by Henderson (1950). We propose a parametric bootstrap MSE estimator of the predictor. We prove its asymptotic unbiasedness and derive the order of its bias. Considerations are supported by Monte Carlo simulation analyses to compare its accuracy (not only the bias) with other MSE estimators, including jackknife and weighted jackknife MSE estimators that we adapt for the considered predictor.

Key words: Empirical best linear unbiased predictor; model approach in survey sampling; parametric bootstrap; properties of MSE estimators; small area estimation.

1. Introduction

To estimate or to predict subpopulation characteristics with small or even zero sample sizes, small area estimation methods are used. In the model approach, empirical (estimated) versions of the best linear unbiased predictor (EBLUP) proposed by Henderson (1950) are widely studied. The first approximation (based on the Taylor's expansion) of the MSE of the predictor is proposed by Kackar and Harville (1984), but they do not study the order of the approximation and the order of the bias of their MSE estimator. These problems are discussed by Prasad and Rao (1990), although in cases where estimators of model parameters are unbiased. Datta and Lahiri (2000) generalize the idea for biased estimators of model parameters including maximum likelihood and restricted maximum likelihood estimators as special cases. More general model and different assumptions leading to different asymptotic results are studied by Das et al. (2004). In all of the above mentioned papers, the additional term of the MSE, resulting from the estimation of model parameters, is derived based on the Taylor's expansion. However, alternatively the jackknife method, proposed by Jiang et al. (2002), or the weighted jackknife method, see Chen and Lahiri (2002, 2003), can be used as well. A similar solution, but using the parametric bootstrap method, is presented in Butar and Lahiri (2003). The parametric bootstrap MSE estimator can also be defined in a different

[1] University of Economics in Katowice, Department of Statistics, Econometrics and Mathematics, 50, 1 Maja Street, 40-287 Katowice, Poland. Email: tomasz.zadlo@ue.katowice.pl

way and under more general models as shown by González-Manteiga et al. (2007), and González-Manteiga et al. (2007, 2008).

However, the empirical version of the best linear unbiased predictor proposed by Henderson (1950) is not the predictor of the linear combination of the variable of interest (that we would like to predict). This predictor can also be used to make predictions for the unsampled part of the population, as described for example, in Rao and Molina (2015, 7.1.3). Even in this case, they are not generally the best linear unbiased predictors of the linear combination of the variable of interest. The additional condition is required, as we show in Section 3.

In this case, as well as under some additional assumptions presented in Section 3, the resulting predictor becomes the empirical best linear unbiased predictor (EBLUP) of the linear combination of the variable of interest.

A more general predictor is the empirical version of the best linear unbiased predictor proposed by Royall (1976), but it is very rarely considered in the small area estimation literature, for example, Rao and Molina (2015, 178) only mentioned it once. We compare it analytically with the empirical version of the best linear unbiased predictor proposed by Henderson (1950). The aim of the article is to propose an asymptotically unbiased parametric bootstrap MSE estimator of the predictor and derive the order of its bias, generalizing the results of Butar and Lahiri (2003) for:

- a more general predictor (see Section 3 and Remark 1 in Section 4),
- and a more general model, taking into account the problem of changes of domain affiliations of population elements and covering many models considered in small area estimation.

Furthermore, our proposal of the MSE estimator will have the advantage that the additional component of the MSE resulting from the estimation of model parameters will not have to be derived, unlike in the case of Taylor's expansion, considered, for example, by Żądło (2009).

In Section 2, we will present our proposal of the superpopulation model, which is a longitudinal linear mixed model with the block-diagonal covariance matrix. It covers many models known from small area estimation literature and allows taking into account possible changes of population and subpopulations in time. In Section 3, we will show that the BLUP proposed by Royall (1976) is a generalization of the BLUP proposed by Henderson (1950) together with a condition when they are equivalent. The dependence between their MSE components will be also presented. In Section 4, we will study the empirical version of the BLUP proposed by Royall (1976) and propose an estimator of its MSE based on the parametric bootstrap method. We will also prove that the order of the bias of the proposed bootstrap MSE estimator is $o(K^{-1})$, where K is the number of blocks in the covariance matrix of the variable of interest. Considerations are supported by simulation studies based on real data.

2. Longitudinal Superpopulation Model

We will introduce a superpopulation model which belongs to the class of linear mixed models with a block-diagonal variance-covariance matrix of the variable of interest.

For further considerations the diagonal structure will be crucial, but we will also show that our model is (in our opinion) very flexible in the sense that it covers different types of longitudinal data even in cases of population and subpopulation changes in time. We will also present many longitudinal models used for small area estimation purposes as special cases of our proposal. In the next two sections, different predictors and their MSEs will be studied under the model.

Longitudinal data for periods $t = 1,..., M$ are considered, where the number M may include future periods. In the period t the population of size N_t is denoted by Ω_t. The population in the period t is divided into D disjoint subpopulations (domains) Ω_{dt} of size N_{dt}, where $d = 1,..., D$. Let the set of population elements for which observations are available in the period t be denoted by s_t and its size by n_t. The set of subpopulation elements for which observations are available in the period t is denoted by s_{dt} and its size by n_{dt}. Let $\Omega_{rdt} = \Omega_{dt} \backslash s_{dt}, N_{rdt} = N_{dt} - n_{dt}$, $\cup_{t=1}^{M} \Omega_t = \Omega$, $\bar{\bar{\Omega}} = N$, $\cup_{t=1}^{M} \Omega_{dt} = \Omega_d$, $\bar{\bar{\Omega}}_d = N_d$, $\cup_{t=1}^{M} \Omega_{rdt} = \Omega_{rd}$, $\bar{\bar{\Omega}}_{rd} = N_{rd}$, $\cup_{t=1}^{M} s_t = s$, $\bar{\bar{s}} = n$, $\cup_{t=1}^{M} s_{dt} = s_d$, $\bar{\bar{s}}_d = n_d$, If the period t is the future period, then $s_t = s_{dt} = \emptyset$, $n_t = n_{dt} = 0$, $\Omega_{rdt} = \Omega_{dt}$ and $N_{rdt} = N_{dt}$.

Let the vector of random variables of interest in M periods, denoted by $\mathbf{Y}$, be divided into K subvectors denoted by $\mathbf{Y}_k$, where $k = 1, 2,..., K$. The division can be made according to any rule, including but not limited to the division based on subsets Ω_t ($t = 1,..., M$ and hence $K = M$), Ω_d ($d = 1,..., D$ and hence $K = D$) or Ω_{dt} ($t = 1,..., M$, $d = 1, 2,..., D$ and hence $K = D \times M$). We assume that population data obey the following model:

$$\begin{cases} \mathbf{Y}_k = \mathbf{X}_k\beta + \mathbf{Z}_k\mathbf{v}_k + \mathbf{e}_k \\ E(\mathbf{e}_k) = 0 \\ E(\mathbf{v}_k) = 0 \\ D^2\begin{bmatrix} \mathbf{v}_k \\ \mathbf{e}_k \end{bmatrix} = \begin{bmatrix} \mathbf{G}_k(\boldsymbol{\delta}) & 0 \\ 0 & \mathbf{R}_k(\boldsymbol{\delta}) \end{bmatrix} \end{cases}, \tag{1}$$

where $\mathbf{Y}_k$ of size $N_k \times 1$ (where $k = 1, 2,..., K$) are assumed to be independent, $\mathbf{X}_k$ and $\mathbf{Z}_k$ are known matrices of auxiliary variables of sizes $N_k \times p$ and $N_k \times r_k$, respectively, and of full ranks, $\boldsymbol{\delta}$ is a vector of q unknown parameters, β is a vector of p unknown parameters, random vectors $\mathbf{e}_k$ and $\mathbf{v}_k$ of sizes $N_k \times 1$ and $r_k \times 1$, respectively, are independent and the symbol $D^2(.)$ denotes the variance-covariance matrix. Hence, the covariance matrix of $\mathbf{Y}_k$ denoted by $D^2(\mathbf{Y}_k) = \mathbf{V}_k(\boldsymbol{\delta})$ is given by:

$$\mathbf{V}_k(\boldsymbol{\delta}) = \mathbf{Z}_k\mathbf{G}_k(\boldsymbol{\delta})\mathbf{Z}_k^T + \mathbf{R}_k(\boldsymbol{\delta}). \tag{2}$$

If we assume, without loss of generality, that first n_k elements of $\mathbf{Y}_k$ are observed then matrices in Equation (1) can be decomposed as follows: where $\mathbf{Y}_k = \begin{bmatrix} \mathbf{Y}_{sk}^T & \mathbf{Y}_{rk}^T \end{bmatrix}^T$, $\mathbf{X}_k = \begin{bmatrix} \mathbf{X}_{sk}^T & \mathbf{X}_{rk}^T \end{bmatrix}^T$, $\mathbf{Z}_k = \begin{bmatrix} \mathbf{Z}_{sk}^T & \mathbf{Z}_{rk}^T \end{bmatrix}^T$, $\mathbf{e}_k = \begin{bmatrix} \mathbf{e}_{sk}^T & \mathbf{e}_{rk}^T \end{bmatrix}^T$, $\mathbf{Y}_{sk}$, $\mathbf{Y}_{rk}$, $\mathbf{X}_{sk}$, $\mathbf{X}_{rk}$, $\mathbf{Z}_{sk}$, $\mathbf{Z}_{rk}$, $\mathbf{e}_{sk}$, $\mathbf{e}_{rk}$ are of sizes $n_k \times 1$, $(N_k - n_k) \times 1$, $n_k \times p$, $(N_k - n_k) \times p$, $n_k \times r_k$, $(N_k - n_k) \times r_k$, $n_k \times 1$, $(N_k - n_k) \times 1$, respectively. In this case, covariance matrices in Equation (2) are given by: $\mathbf{V}_k = \begin{bmatrix} \mathbf{V}_{kss} & \mathbf{V}_{ksr} \\ \mathbf{V}_{krs} & \mathbf{V}_{krr} \end{bmatrix}$ and $\mathbf{R}_k = \begin{bmatrix} \mathbf{R}_{kss} & \mathbf{R}_{ksr} \\ \mathbf{R}_{krs} & \mathbf{R}_{krr} \end{bmatrix}$, where matrices $\mathbf{V}_{kss}$ and $\mathbf{R}_{kss}$ are $n_k \times n_k$, $\mathbf{V}_{krr}$ and $\mathbf{R}_{krr}$ are $(N_k - n_k) \times (N_k - n_k)$, $\mathbf{V}_{ksr}$ and $\mathbf{R}_{ksr}$ are $n_k \times (N_k - n_k)$, $\mathbf{V}_{krs} = \mathbf{V}_{ksr}^T$ and $\mathbf{R}_{krs} = \mathbf{R}_{ksr}^T$.

Model (1) can be written as follows:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{v} + \mathbf{e}, \tag{3}$$

where $\mathbf{Y} = col_{1\le k\le K}(\mathbf{Y}_k)$, $\mathbf{X} = col_{1\le k\le K}(\mathbf{X}_k)$, $\mathbf{Z} = diag_{1\le k\le K}(\mathbf{Z}_k)$, $\mathbf{v} = col_{1\le k\le K}(\mathbf{v}_k)$, $\mathbf{e} = col_{1\le k\le K}(\mathbf{e}_k)$. Hence, the covariance matrix of $\mathbf{Y}$ denoted by $D^2(\mathbf{Y}) = \mathbf{V}(\boldsymbol{\delta})$ is given by:

$$\mathbf{V}(\boldsymbol{\delta}) = \mathbf{Z}\mathbf{G}(\boldsymbol{\delta})\mathbf{Z}^T + \mathbf{R}(\boldsymbol{\delta}), \tag{4}$$

where $\mathbf{G} = diag_{1\le k\le K}(\mathbf{G}_k)$ and $\mathbf{R} = diag_{1\le k\le K}(\mathbf{R}_k)$.

Model (1) is a very general model covering many special cases, including cases presented below, in which changes of the population and subpopulations are taken into account. The assumption that one population element may change its domain (or group) affiliation in time is very important in longitudinal surveys. For example, let us consider the population of households and the division of the population into domains made according to household size. In this case, we should assume that some households can change their sizes in time and hence their domain affiliation. If some human population is under the study, one may be interested in its characteristics for subpopulations defined according to some social or economic criteria (e.g., the job position). In the case of business surveys, the population of firms may be divided into subpopulations according to some economic or financial criteria, which may imply even stronger changes in the division of the population in time than in the case of human populations.

Let M_{id} denotes the number of periods when the ith population element belongs to the dth domain (it may include future periods). Values of the variable of interest are realizations of random variables Y_{idj} for the ith population element that belongs to the dth domain in the period t_{ij}, where $i = 1, 2,..., N$, $j = 1, 2,..., M_{id}$, $d = 1, 2,..., D$. The vector $\mathbf{Y}_{id} = [Y_{idj}]_{M_{id}} \times 1$ is called the profile. We have defined $\Omega = \cup_{t=1}^{M} \Omega_t$, where Ω_t is the population in the period t. Now, let us assume that profiles in Ω are divided into K groups Ω_k for which where the number of random variables within each group equals N_k, where $k = 1, 2,..., K$. Grouping profiles means that groups always have a longitudinal character in this special case of Model (1). We consider the following four cases:

- C1 - the additional division of the profiles is not taken into account and hence, the number of groups of profiles equals the number of profiles ($K = N$),
- C2 - domains Ω_d ($d = 1, 2,..., D$) are unions of groups of profiles Ω_k ($K > D$ and $K \neq N$),
- C3 - groups of profiles are domains Ω_d ($d = 1, 2,..., D$) and hence $K = D$,
- C4 - sets Ω_k ($k = 1, 2,..., K$) are unions of domains Ω_d ($K < D$).

Now we can define the profile more precisely, including the information on the division of the population into K groups of profiles. The vector of random variables for the ith population element in different periods will be denoted by $\mathbf{Y}_{idk} = \left[Y_{idjk}\right]_{M_{idk}\times 1,}$ where M_{idk} is the number of periods when the ith element belongs to the dth domain and the kth group of profiles. Let the number of profiles within the kth group be denoted by M_k. Hence, the number of random variables within Ω_k equals $N_k = \sum_{i=1}^{M_k} M_{idk}$. In the cases C1 and C3, the additional subscript k is not necessary because i and d explicitly define k. Let m_{idk} be the number of periods when the ith population element (which belongs to the dth domain

and the kth group of profiles) is observed. The vector $\mathbf{Y}_{sidk} = \left[Y_{idkj}\right]_{m_{idk}\times 1}$, will be called the sample profile. Let the number of profiles observed in the sample in the kth group be denoted by m_k. Hence, the number of sample observations from Ω_k equals $n_k = \sum_{i=1}^{m_k} m_{idk}$. Let the vector $\mathbf{Y}_{ridk} = \left[Y_{idkj}\right]_{M_{ridk}\times 1}$, where $M_{ridk} = M_{idk} - m_{idk}$, be the profile of random variables with non-observed realizations (including out of sample and future values). This notation allows the inclusion of possibilities of changes of the population in time, changes of subpopulations in time and even changes of domains or group affiliations of population elements in time in the considerations.

Let us consider the cases in which random variables for one population element observed in different periods form more than one profile. It is possible in the following cases:

- for C1: a change of the domain affiliation,
- for C2: a change of the group affiliation within a domain or a change of the domain affiliation,
- for C3: a change of the domain affiliation equivalent to a change of the group affiliation,
- for C4: the change of the domain affiliation within the group or the change of the group and the domain affiliation at the same time.

It must be pointed out that the period of time can be denoted by one out of two indexes: t (where $t = 1, 2,..., M$) - to distinguish between different periods of the longitudinal data, and j (where $j = 1, 2,..., M_{id}$) - to distinguish between different periods within profiles.

Model (1) covers many unit-level longitudinal models with block-diagonal covariance matrix considered in the literature. Case C1 of the model includes, for example, the model with independent profile specific random effects considered by Verbeke and Molenbergh (2000, 20). Random regression coefficient models considered by Hobza and Morales (2013) and the multilevel model studied by Moura and Holt (1999), both assumed for one period, and also heteroscedastic models with domain specific and domain and domain-and-time specific random effects considered by Morales and Santamaría (2019) are special cases of case C3. The model with two random effects (domain specific and profile specific) considered by Stukel and Rao (1999) and Nissinen (2009, 22) and the model with two random effects (domain specific and domain-and-time specific) considered by Molina et al. (2010, 143–180) are covered by the case C3 of our model as well. Case C3 also covers models considered by Żądło (2014, 2015a) with profile-specific random effects spatially correlated within domains. Case C4 of the model includes the model with profile-specific random effects spatially correlated within groups of domains proposed by Żądło (2015b). What is more, longitudinal area-level models considered by Rao and Yu (1994) and Marhuenda et al. (2013) can also be written as a Case C3 of our model. However, not all special cases of Model (1) are covered by Cases C1-C4 due to assumed non-longitudinal character of groups including the longitudinal model with domain-and-time specific random effects (autocorrelated in time) considered by Saei and Chambers (2003,13).

It must be pointed out that Model (1) does not cover linear mixed models without block-diagonal covariance matrix, see models studied by, for example Fabrizi et al. (2007, 189), Pagliarell and Salvatore (2016, 232–235), D'Aló et al. (2017), and nonlinear mixed

models studied by, for example Hobza et al. (2018). The overview of different small area models can be found in Jiang and Lahiri (2006).

3. Best Linear Unbiased Predictors

Under the linear mixed model presented in the previous section, we will compare analytically different predictors and their MSEs, including a predictor that is widely discussed in the literature, proposed by Henderson (1950) and a more general predictor presented by Royall (1976).

We study the problem of prediction of a linear combination $\theta = \gamma^T \mathbf{Y}$. For example, if we are interested in the prediction of the dth subpopulation (domain) total in the period t then the kth element of the vector γ equals 1 for $k \in \Omega_{dt}$ and 0 otherwise. Under Model (3):

$$\theta = \gamma^T \mathbf{Y} = \gamma^T \mathbf{X}\beta + \gamma^T \mathbf{Z}\mathbf{v} + \gamma^T \mathbf{e}. \tag{5a}$$

Let us present the results obtained by Henderson (1950) for surveys conducted in one period by changing the sizes of matrices in his model to cover the case of the longitudinal Model (1). He considers the problem of prediction of:

$$\theta^s = \mathbf{l}^T \beta + \mathbf{m}^T \mathbf{v}. \tag{5b}$$

For $\mathbf{l}^T = \gamma^T \mathbf{X}$ and $\mathbf{m}^T = \gamma^T \mathbf{Z}$ from Equations (5a) and (5b) we obtain:

$$\theta = \theta^s + \gamma^T \mathbf{e}. \tag{5c}$$

Theorem 1. (see Henderson 1950). Assume that sample data obey the following assumptions:

$$\begin{cases} \mathbf{Y}_s = \mathbf{X}_s\beta + \mathbf{Z}\mathbf{v} + \mathbf{e}_s \\ E(\mathbf{e}_s) = 0 \\ E(\mathbf{v}) = 0 \\ D^2 \begin{bmatrix} \mathbf{v} \\ \mathbf{e}_s \end{bmatrix} = \begin{bmatrix} \mathbf{G}(\boldsymbol{\delta}) & 0 \\ 0 & \mathbf{R}_{ss}(\boldsymbol{\delta}) \end{bmatrix} \end{cases} \tag{6}$$

Among linear, model-unbiased predictors $\hat{\theta}^s = \mathrm{a}^T \mathbf{Y}_s + \mathbf{b}$ of linear combination of β and the realization of $\mathbf{v}$ given by $\theta^s = \mathbf{l}^T \beta + \mathbf{m}^T \mathbf{v}$ (for specified vectors, $\mathbf{l}$ and $\mathbf{m}$, of constants) the MSE is minimized by:

$$\hat{\theta}^s_{BLUP} = \mathbf{1}^T \tilde{\beta}(\boldsymbol{\delta}) + \mathbf{m}^T \tilde{\mathbf{v}}(\boldsymbol{\delta}), \tag{7a}$$

where

$$\tilde{\beta}(\boldsymbol{\delta}) = \left(\mathbf{X}_s^T \mathbf{V}_{ss}^{-1}(\boldsymbol{\delta}) \mathbf{X}_s\right)^{-1} \mathbf{X}_s^T \mathbf{V}_{ss}^{-1}(\boldsymbol{\delta}) \mathbf{Y}_s, \tag{7b}$$

$$\tilde{\mathbf{v}}(\boldsymbol{\delta}) = \mathbf{G}(\boldsymbol{\delta}) \mathbf{Z}_s^T \mathbf{V}_{ss}^{-1}(\boldsymbol{\delta}) \left(\mathbf{Y}_s - \mathbf{X}_s \tilde{\beta}(\boldsymbol{\delta})\right). \tag{7c}$$

The MSE of $\hat{\theta}^s_{BLUP}$ is given by

$$MSE\left(\hat{\theta}^s_{BLUP}\right) = Var\left(\hat{\theta}^s_{BLUP} - \theta^s\right) = g^s_1(\boldsymbol{\delta}) + g^s_2(\boldsymbol{\delta}), \tag{7d}$$

where

$$g^s_1(\boldsymbol{\delta}) = \mathbf{m}^T\left(\mathbf{G}(\boldsymbol{\delta}) - \mathbf{G}(\boldsymbol{\delta})\mathbf{Z}^T_s\mathbf{V}^{-1}_{ss}(\boldsymbol{\delta})\mathbf{Z}_s\mathbf{G}(\boldsymbol{\delta})\right)\mathbf{m}, \tag{7e}$$

$$\begin{aligned} g^s_1(\boldsymbol{\delta}) = &\left(\mathbf{1}^T - \mathbf{m}^T\mathbf{G}(\boldsymbol{\delta})\mathbf{Z}^T_s\mathbf{V}^{-1}_{ss}(\boldsymbol{\delta})\mathbf{X}_s\right)\left(\mathbf{X}^T_s\mathbf{V}^{-1}_{ss}(\boldsymbol{\delta})\mathbf{X}_s\right)^{-1}\times \\ &\times\left(\mathbf{1}^T - \mathbf{m}^T\mathbf{G}(\boldsymbol{\delta})\mathbf{Z}^T_s\mathbf{V}^{-1}_{ss}(\boldsymbol{\delta})\mathbf{X}_s\right)^T. \end{aligned} \tag{7f}$$

The proof of the theorem (for surveys conducted in one period) is presented in detail in, for example, Rao and Molina (2015, 119–120).

Because in Theorem 1 the problem of prediction of Equation (5b) instead of Equation (5c) is considered, in small area estimation literature, see, for example, Rao and Molina (2015, 178–179), the problem of prediction of Equation (5c) using predictor (7a) is studied. We assume longitudinal Model (1) and study the problem of prediction of Equation (5c), but in a more general framework than in Rao and Molina (2015, 178–179), where the crucial difference is the lack of the independence assumption of random components.

Let us consider the problem of prediction of

$$\theta = \boldsymbol{\gamma}^T\mathbf{Y} = \boldsymbol{\gamma}^T_s\mathbf{Y}_s + \boldsymbol{\gamma}^T_r\mathbf{Y}_r = \boldsymbol{\gamma}^T_s\mathbf{Y}_s + \theta_r, \tag{8}$$

where $\boldsymbol{\gamma} = \left[\boldsymbol{\gamma}^T_s\,\boldsymbol{\gamma}^T_r\right]^T$, $\theta_r = \boldsymbol{\gamma}^T_r\mathrm{Y}_r$ and the realization of $\boldsymbol{\gamma}^T_s\mathrm{Y}_s$ is known. Because realization of $\boldsymbol{\gamma}^T_s\mathrm{Y}_s$ is known, the problem of prediction of Equation (8) is reduced to the problem of prediction of:

$$\theta_r = \boldsymbol{\gamma}^T_r\mathbf{Y}_r = \boldsymbol{\gamma}^T_r\mathbf{X}_r\boldsymbol{\beta} + \boldsymbol{\gamma}^T_r\mathbf{Z}_r\mathbf{v} + \boldsymbol{\gamma}^T_r\mathbf{e}_r = \theta^s_r + \boldsymbol{\gamma}^T_r\mathbf{e}_r, \tag{9}$$

where

$$\theta^s_r = \boldsymbol{\gamma}^T_r\mathbf{X}_r\boldsymbol{\beta} + \boldsymbol{\gamma}^T_r\mathbf{Z}_r\mathbf{v}. \tag{10}$$

Based on Theorem 1 the BLUP of Equation (10) is given by:

$$\hat{\theta}^s_{BLUPr} = \mathbf{1}^T\tilde{\boldsymbol{\beta}}(\boldsymbol{\delta}) + \mathbf{m}^T\tilde{\mathbf{v}}(\boldsymbol{\delta}), \tag{11}$$

where $\mathbf{1}^T = \boldsymbol{\gamma}^T_r\mathbf{X}_r, \mathbf{m}^T = \boldsymbol{\gamma}^T_r\mathbf{Z}_r, \tilde{\boldsymbol{\beta}}(\boldsymbol{\delta})$ and $\tilde{\mathbf{v}}(\boldsymbol{\delta})$ are given by Equations (7b) and (7c), respectively.

Finally, we obtain the following predictor:

$$\begin{aligned} &\hat{\theta} = \boldsymbol{\gamma}^T_s\mathbf{Y}_s + \hat{\theta}^s_{BLUPr} = \boldsymbol{\gamma}^T_s\mathbf{Y}_s + \\ &+\boldsymbol{\gamma}^T_r\left[\mathbf{X}_r\tilde{\boldsymbol{\beta}}(\boldsymbol{\delta}) + \left(\mathbf{Z}_r\mathbf{G}(\boldsymbol{\delta})\mathbf{Z}^T_s\right)\mathbf{V}^{-1}_{ss}(\boldsymbol{\delta})\left(\mathbf{Y}_s - \mathbf{X}_s\tilde{\boldsymbol{\beta}}(\boldsymbol{\delta})\right)\right] \end{aligned} \tag{12a}$$

of Equation (8), which generally is not the BLUP of Equation (8). Later in this section, we will show that if $\mathbf{R}_{sr} = 0$, then the predictor given by Equation (12a) becomes the BLUP of Equation (8).

The MSE of the predictor (12a) of Equation (8) is given by

$$\begin{aligned} MSE(\hat{\theta}) &= E\left(\gamma_s^T \mathbf{Y}_s + \hat{\theta}^s_{BLUPr} - \gamma^T \mathbf{Y}\right)^2 = \\ &= E\left(\hat{\theta}^s_{BLUPr} - \gamma_r^T \mathbf{Y}_r\right)^2 = \\ &= E\left(\hat{\theta}^s_{BLUPr} - \theta^s_r - \gamma_r^T \mathbf{e}_r\right)^2 = \\ &= MSE\left(\hat{\theta}^s_{BLUPr}\right) + \\ &\quad + g_4^s(\boldsymbol{\delta}) - 2Cov\left(\left(\hat{\theta}^s_{BLUPr} - \theta^s_r\right), \gamma_r^T \mathbf{e}_r\right), \end{aligned} \tag{12b}$$

where $MSE\left(\hat{\theta}^s_{BLUPr}\right)$ is given by Equation (7d) (where $\mathbf{1}^T = \gamma_r^T \mathbf{X}_r$ and $\mathbf{m}^T = \gamma_r^T \mathbf{Z}_r$), $g_4^s(\boldsymbol{\delta}) = \gamma_r^T \mathbf{R}_{rr}(\delta)\gamma_r$ and

$$\begin{aligned} &Cov\left(\left(\hat{\theta}^s_{BLUPr} - \theta^s_r\right), \gamma_r^T \mathbf{e}_r\right) = \\ &= \gamma_r^T \mathbf{X}_r \left(\mathbf{X}_s^T \mathbf{V}_{ss}^{-1}(\boldsymbol{\delta}) \mathbf{X}_s\right)^{-1} \mathbf{X}_s^T \mathbf{V}_{ss}^{-1} \mathbf{R}_{st}(\boldsymbol{\delta})\gamma_r + \\ &+ \gamma_r^T \mathbf{Z}_r \mathbf{G}(\boldsymbol{\delta}) \mathbf{Z}_s^T \mathbf{V}_{ss}^{-1}(\boldsymbol{\delta}) \times \\ &\times \left(\mathbf{I} - \mathbf{X}_s \left(\mathbf{X}_s^T \mathbf{V}_{ss}^{-1}(\boldsymbol{\delta}) \mathbf{X}_s\right)^{-1} \mathbf{X}_s^T \mathbf{V}_{ss}^{-1}(\boldsymbol{\delta})\right) \mathbf{R}_{sr}(\boldsymbol{\delta})\gamma_r. \end{aligned} \tag{12c}$$

Firstly, we assume that $\mathrm{R}_{sr} = 0$. Then, $Cov\left(\left(\hat{\theta}^s_{BLUPr} - \theta^s_r\right), \gamma_r^T \mathbf{e}_r\right) = 0$ and Equation (12b) simplifies to

$$\begin{aligned} MSE(\hat{\theta}) &= MSE\left(\hat{\theta}^s_{BLUPr}\right) + g_4^s(\boldsymbol{\delta}) = \\ &= g_1^s(\boldsymbol{\delta}) + g_2^s(\boldsymbol{\delta}) + g_4^s(\boldsymbol{\delta}) \end{aligned} \tag{12d}$$

and the predictor (12a) is the BLUP of Equation (8) (which will be shown later in this section).

Royall (1976) for surveys conducted in one period derived a more general predictor than predictors (7a) and (12a). In the following consideration we change the sizes of matrices in his theorem to cover the following longitudinal model:

$$\begin{cases} E(\mathbf{Y}) = \mathbf{X}\beta \\ D^2(\mathbf{Y}) = \mathbf{V} \end{cases}, \tag{13}$$

where matrices in Equation (13) were defined in Section 1.

Theorem 2. (compare Royall 1976). Assume that the population data obey the longitudinal general linear model (see Equation (13)). Among linear, model-unbiased predictors $\hat{\theta} = g^T \mathbf{Y}_s$ of linear combination of random variables $\theta = \gamma^T \mathbf{Y}$, the MSE is

minimized by:

$$\hat{\theta}_{BLUP} = \gamma_s^T \mathbf{Y}_s + \gamma_r^T \left[\mathbf{X}_r \tilde{\beta}(\boldsymbol{\delta}) + \mathbf{V}_{rs}(\boldsymbol{\delta}) \mathbf{V}_{ss}^{-1}(\boldsymbol{\delta}) \left(\mathbf{Y}_s - \mathbf{X}_s \tilde{\beta}(\boldsymbol{\delta}) \right) \right], \tag{14a}$$

where $\tilde{\boldsymbol{\beta}}(\boldsymbol{\delta})$ is given by (7b).

The MSE of $\hat{\theta}_{BLUP}$ is given by

$$MSE\left(\hat{\theta}_{BLUP}\right) = g_1(\boldsymbol{\delta}) + g_2(\boldsymbol{\delta}), \tag{14b}$$

where

$$g_1(\boldsymbol{\delta}) = \gamma_r^T \left(\mathbf{V}_{rr}(\boldsymbol{\delta}) - \mathbf{V}_{rs}(\boldsymbol{\delta}) \mathbf{V}_{ss}^{-1}(\boldsymbol{\delta}) \mathbf{V}_{sr}(\boldsymbol{\delta}) \right) \gamma_r, \tag{14c}$$

$$\begin{aligned} g_2(\boldsymbol{\delta}) = {} & \gamma_r^T \left(\mathbf{X}_r - \mathbf{V}_{rs}(\boldsymbol{\delta}) \mathbf{V}_{ss}^{-1}(\boldsymbol{\delta}) \mathbf{X}_s \right) \\ & \times \left(\mathbf{X}_s^T \mathbf{V}_{ss}^{-1}(\boldsymbol{\delta}) \mathbf{X}_s \right)^{-1} \times \left(\mathbf{X}_r - \mathbf{V}_{rs}(\boldsymbol{\delta}) \mathbf{V}_{\mathrm{ss}}^{-1}(\boldsymbol{\delta}) \mathbf{X}_s \right)^{\mathrm{T}} \gamma_r. \end{aligned} \tag{14d}$$

The proof of the theorem is presented in details by, for example, Valliant et al. (2000, 29–30).

Firstly, let us compare predictor (14a) proposed by Royall (1976) with predictor (7a) proposed by Henderson (1950). Predictor (14a) can be treated as a generalization of the predictor (7a) because (i) the general linear Model (13) covers the assumptions of the general linear mixed Model (6), (ii) the predictor (14a) is the predictor of θ, while the predictor (7a) is the predictor of θ^s (where (5c) holds) and (iii) $\hat{\theta}^s = \mathrm{a}^T \mathrm{Y}_s + b$ considered in Theorem 1 is of the same form as $\hat{\theta} = \mathrm{g}^T \mathrm{Y}_\mathrm{s}$ considered in Theorem 2, because $b = 0$ under unbiasedness of the predictor.

Secondly, let us compare predictors of θ given by Equation (14a), proposed by Royall (1976), and given by Equation (12a). Although (14a) is derived under the general linear Model (13), we consider its formula under the special case of the model - the general linear mixed Model (1). Hence, the predictor (14a) simplifies to:

$$\begin{aligned} & \hat{\theta}_{BLUP} = \gamma_s^T \mathbf{Y}_s + \gamma_r^T \mathbf{X}_r \tilde{\beta}(\boldsymbol{\delta}) + \\ & + \gamma_r^T \left(\mathbf{R}_{rs}(\boldsymbol{\delta}) + \mathbf{Z}_r \mathbf{G}(\boldsymbol{\delta}) \mathbf{Z}_s^T \right) \mathbf{V}_{ss}^{-1}(\boldsymbol{\delta}) \left(\mathbf{Y}_s - \mathbf{X}_s \tilde{\beta}(\boldsymbol{\delta}) \right). \end{aligned} \tag{15}$$

If $\mathbf{R}_{rs} = 0$ then predictors (15) are (12a) identical and their MSEs (given by Equations (14b) and (12d), respectively) are equal. Being more precise, if $\mathbf{R}_{rs} = 0$, then

$$g_1(\boldsymbol{\delta}) = g_1^s(\boldsymbol{\delta}) + g_4^s(\boldsymbol{\delta}) \tag{16}$$

and

$$g_2(\boldsymbol{\delta}) = g_2^s(\boldsymbol{\delta}). \tag{17}$$

To sum up, the predictor (12a) is the BLUP of θ only for models where $\mathbf{R}_{rs} = 0$.

4. Empirical Best Linear Unbiased Predictors

In the previous section, we assumed that model parameters are known. Now we will take into account additional variability resulting from their estimation and its influence on the MSE and MSE estimation.

Let us start our considerations from the predictor proposed by Henderson (1950). The Best Linear Unbiased Predictor (BLUP) (7a) depends on the variance parameters $\boldsymbol{\delta}$ that are unknown in practical applications. Replacing $\boldsymbol{\delta}$ by an estimator $\hat{\hat{\boldsymbol{\delta}}}$, we obtain the two-stage predictor $\hat{\theta}^s_{EBLUP} = \hat{\theta}^s_{BLUP}(\hat{\boldsymbol{\delta}})$ called empirical (estimated) best linear unbiased predictor EBLUP. In the simulation we will use the Restricted Maximum Likelihood (REML) method, known as robust on nonnormality (Jiang 1996) to estimate $\boldsymbol{\delta}$. Under some weak assumptions presented by Kackar and Harville (1981):

(i) an estimator of the vector of model parameters is any even, translation-invariant estimator, for example the REML estimator,
(ii) the distributions of $\mathbf{v}$ and $\mathbf{e}_s$ are both symmetric around 0,
(iii) the expectation of the predictor is finite, the EBLUP remains unbiased.

The MSE of EBLUP is greater than the MSE of BLUP due to the estimation of the model parameters. The problem of estimating accuracy of the EBLUP is widely discussed in the literature and results discussed below can be directly used for our case just by changing the sizes of the matrices to cover the longitudinal Model (1). The first paper where the problem of the approximate formula (based on the Taylor expansion) of the MSE of the EBLUP is studied and its estimator is proposed is Kackar and Harville (1984). The authors do not study the orders of the neglected terms in the approximation of the MSE and the order of the bias of the MSE estimator. The pioneering results are presented by Prasad and Rao (1990). They assume, inter alia, block-diagonal variance-covariance matrix, normality of random effects and random components and unbiasedness of $\boldsymbol{\delta}$ which allows obtaining the order of MSE approximation and proving the asymptotically unbiasedness of their MSE estimator. The problem is also studied in detail by Datta and Lahiri (2000), but for a larger class of estimators $\boldsymbol{\delta}$ (including maximum likelihood and restricted maximum likelihood estimators), where unbiasedness is not required. The more general mixed model is studied by Das et al. (2004), who, based on different assumptions obtained different asymptotic results. Jackknife and weighted jackknife estimators of MSE of EBLUP are proposed by Jiang et al. (2002), Chen and Lahiri (2002, 2003). The parametric bootstrap method is used to estimate MSE of EBLUP by, inter alia, Butar and Lahiri (2003), González-Manteiga et al. (2007, 2008) As shown, for example, by Schmid and Münnich (2014), the parametric bootstrap method of MSE estimation is preferable in the case of different predictors as well.

If $\boldsymbol{\delta}$ in the formula of the predictor (14a) proposed by Royall (1976) is replaced by an estimator $\hat{\hat{\boldsymbol{\delta}}}$, we obtain $\hat{\theta}_{EBLUP} = \hat{\theta}_{BLUP}(\hat{\boldsymbol{\delta}})$ - the empirical version of the predictor. It remains unbiased under some weak assumptions presented by Żądło (2004) ((i) and (iii) the same as mentioned above in the case of empirical version of the predictor proposed by Henderson (1950), in assumption (ii) symmetric distribution around zero of $\mathbf{e}$ instead of $\mathbf{e}_s$ is assumed). For the empirical version of the predictor studied by Royall (1976), the MSE estimator using Taylor's series expansion was proposed by Żądło (2009), who

generalized the results of Datta and Lahiri (2000) obtained for the empirical version of the predictor considered by Henderson (1950). In this section, we will propose a parametric bootstrap MSE estimator by a generalization of the results presented by Butar and Lahiri (2003). For (i) the more general predictor - we will study the empirical version of the Royall (1976) predictor instead of the empirical version of the Henderson (1950) predictor and (ii) for the more general model (the model proposed in Section 1).

Butar and Lahiri (2003) consider two superpopulation models assumed for sample data. We assume the following superpopulation models, but for the whole population:

- Model 1: Model (1) with an additional assumption of normality of random effects and random components
- Model 2: bootstrap model:

$$\begin{cases} \mathbf{Y}_k^* = \mathbf{X}_k\hat{\boldsymbol{\beta}} + \mathbf{Z}_k\mathbf{v}_k^* + \mathbf{e}_k^* \\ E_*\left(\mathbf{e}_k^*\right) = 0 \\ E_*\left(\mathbf{v}_k^*\right) = 0 \\ D_*^2\begin{bmatrix}\mathbf{v}_k^* \\ \mathbf{e}_k^*\end{bmatrix} = \begin{bmatrix}\mathbf{G}_k\left(\hat{\boldsymbol{\delta}}\right) & 0 \\ 0 & \mathbf{R}_k\left(\hat{\boldsymbol{\delta}}\right)\end{bmatrix} \end{cases}, \tag{18}$$

where $\mathbf{Y}_1^*,\ldots,\mathbf{Y}_k^*,\ldots,\mathbf{Y}_K^*$ are independent, $\mathbf{v}_k^*$ and $\mathbf{e}_k^*$ are generated from multivariate normal distributions, $\hat{\boldsymbol{\delta}}$ is an estimator of $\boldsymbol{\delta}$ that satisfies the regularity conditions (RC) given in the Appendix (Section 7), $\hat{\beta} = \tilde{\beta}(\hat{\boldsymbol{\delta}})$, where $\tilde{\beta}(\boldsymbol{\delta})$ is given by Equation (7b).

We propose the following parametric bootstrap MSE estimator:

$$\begin{aligned} &\widehat{MSE}\left(\hat{\theta}_{EBLUP}\right) = g_1\left(\hat{\boldsymbol{\delta}}\right) + g_2\left(\hat{\boldsymbol{\delta}}\right) + \\ &-E_*\left(g_1\left(\boldsymbol{\delta}^*\right) + g_2\left(\boldsymbol{\delta}^*\right) - \left(g_1\left(\hat{\boldsymbol{\delta}}\right) + g_2\left(\hat{\boldsymbol{\delta}}\right)\right)\right) + \\ &+E_*\left(\hat{\theta}_{EBLUP}\left(\mathrm{Y}_s, \hat{\beta}\left(\boldsymbol{\delta}^*\right), \boldsymbol{\delta}^*\right) - \hat{\theta}_{EBLUP}\left(\mathrm{Y}_s, \hat{\beta}(\boldsymbol{\delta}), \boldsymbol{\delta}\right)\right)^2, \end{aligned} \tag{19}$$

where E_* it is the expectation with respect to Model 2, $\boldsymbol{\delta}^*$ is calculated as $\boldsymbol{\delta}$ but based on $\mathbf{Y}_s^*$ instead of $\mathbf{Y}_s$. The formula of the estimator (19) of the empirical version of Royall (1976) predictor is similar to the MSE estimator of the empirical version of Henderson (1950) predictor proposed by Butar and Lahiri (2003). If $g_1(.)$, $g_2(.)$ and $\hat{\theta}_{EBLUP}$ in (19) are replaced by $g_1^s(\cdot), g_2^s(\cdot)$ and $\hat{\theta}_{EBLUP}^s$ respectively, we obtain the Butar and Lahiri (2003) MSE estimator.

We will prove that the estimator (19) is asymptotically unbiased and that the bias is of order $o(K^{-1})$. The proof will be a generalization of the proof presented by Butar and Lahiri (2003). Butar and Lahiri (2003) propose another MSE estimator which approximates the parametric bootstrap MSE estimator but its bias, Butar and Lahiri (2003), is of a higher order – the same order as the bias of the naive MSE estimator.

Lemma 1. Under Model 1 and the regularity conditions (RC) presented in the Appendix (Section 7), we have

$$E\left(g_1(\boldsymbol{\delta}) - \mathrm{B}_{\boldsymbol{\delta}}^T(\boldsymbol{\delta})\frac{\partial g_1(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}} + g_3(\boldsymbol{\delta})\right) = g_1(\boldsymbol{\delta}) + o\left(K^{-1}\right),$$
$$E\left(g_2(\boldsymbol{\delta})\right) = g_2(\boldsymbol{\delta}) + o\left(K^{-1}\right),$$
$$E\left(g_3(\boldsymbol{\delta})\right) = g_3(\boldsymbol{\delta}) + o\left(K^{-1}\right),$$

where

$$g_3(\boldsymbol{\delta}) = trace\left(\frac{\partial \mathrm{c}^{\mathrm{T}}(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}}\mathbf{V}_{ss}\left(\frac{\partial \mathrm{c}^T(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}}\right)^T \Sigma(\boldsymbol{\delta})\right) \tag{20}$$

and

$$\mathrm{c}^T = \gamma_r^T \mathbf{V}_{rs}\mathbf{V}_{ss}^{-1} = \gamma_r^T\left(\mathbf{R}_{rs} + \mathbf{Z}_r\mathbf{G}\mathbf{Z}_s^T\right)\mathbf{V}_{ss}^{-1},$$

$\Sigma(\boldsymbol{\delta}) = E(\boldsymbol{\delta} - \boldsymbol{\delta})(\boldsymbol{\delta} - \boldsymbol{\delta})^T$, $\mathrm{B}_{\boldsymbol{\delta}}(\boldsymbol{\delta})$ is defined in the regularity condition (f) in the Appendix (Section 7).

Proof. The proof results directly (assuming our regularity conditions, our model and replacing $\mathbf{m}^T\mathbf{G}\mathbf{Z}_s^T\mathbf{V}_{ss}^{-1}$ by $\gamma_r^T\mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}$) from the proof of the theorem of Datta and Lahiri (2000, 624) (called theorem A.2). The difference between Lemma 1 and the theorem presented in Żądło (2009, 110) results from the assumed model (including the number of blocks in the covariance matrix and the sizes of the matrices).

Remark 1. Under normality of $\mathbf{Y}$ the $g_3(\boldsymbol{\delta})$ given by Equation (20) approximates the difference between the MSE of EBLUP and the MSE of BLUP for the predictor considered by Royall (1976) (the predictor presented in Theorem 2) - see Żądło (2009, 107). Applying results presented by Żądło (2009) to RC presented in the Appendix (Section 7), it can be proved that the order of approximation is $o(K^{-1})$. It is worth noting that Equation (20) is the generalization of:

$$g_3^s(\boldsymbol{\delta}) = trace\left(\frac{\partial \mathbf{b}^{\mathrm{T}}(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}}\mathbf{V}_{ss}\left(\frac{\partial \mathbf{b}^T(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}}\right)^T \Sigma(\boldsymbol{\delta})\right), \tag{21}$$

where

$$\mathbf{b}^T = \mathbf{m}^T\mathbf{G}\mathbf{Z}_s^T\mathbf{V}_{ss}^{-1}, \tag{22}$$

presented by Datta and Lahiri (2000), which approximates the difference between the MSE of EBLUP and the MSE of BLUP for the predictor considered by Henderson (1950) (the predictor presented in Theorem 1). Moreover, if we consider the problem of prediction of Equation (10), then Equation (22) is given by:

$$\mathbf{b}^T = \gamma_r^T\mathbf{Z}_r\mathbf{G}\mathbf{Z}_s^T\mathbf{V}_{ss}^{-1}.$$

Then, for models with $R_{rs} = 0$ we obtain

$$g_3^s(\boldsymbol{\delta}) = g_3(\boldsymbol{\delta}), \tag{23}$$

where $g_3(\boldsymbol{\delta}) = g_3^s(\boldsymbol{\delta})$, are given by Equations (21) and (20), respectively.

Lemma 2. Under Model 1, Model 2 and regularity conditions (RC) presented in Appendix (Section 7):

(i) $E_*\left(g_1\left(\boldsymbol{\delta}^*\right)\right) = g_1(\boldsymbol{\delta}) + \mathbf{B}_{\boldsymbol{\delta}}^T(\boldsymbol{\delta})\frac{\partial g_1(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}} - g_3(\boldsymbol{\delta}) + o_p\left(K^{-1}\right)$,
(ii) $E_*\left(g_2\left(\boldsymbol{\delta}^*\right)\right) = g_2(\boldsymbol{\delta}) + o_p\left(K^{-1}\right)$,
(iii) $E_*\left(\hat{\theta}_{EBLUP}\left(\mathbf{Y}_s, \hat{\beta}\left(\boldsymbol{\delta}^*\right), \boldsymbol{\delta}^*\right) - \hat{\theta}_{EBLUP}\left(\mathbf{Y}_s, \hat{\beta}(\boldsymbol{\delta}), \boldsymbol{\delta}\right)\right)^2 = g_4(\boldsymbol{\delta}) + o_p\left(K^{-1}\right)$,
(iv) $E\left(g_4(\boldsymbol{\delta})\right) = g_3(\boldsymbol{\delta}) + o\left(K^{-1}\right)$,

where $g_4(\boldsymbol{\delta}) = trace\left(\frac{\partial c^T(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}}\left(\mathbf{Y}_s - \mathbf{X}_s\hat{\beta}\right)\left(\mathbf{Y}_s - \mathbf{X}_s\hat{\beta}\right)^T\left(\frac{\partial c^T(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}}\right)^T\Sigma(\boldsymbol{\delta})\right)$.

Proof. Parts (i) and (ii) in Lemma 2 follow from (i) and (ii) in Lemma 1 (but under Model 2) using $E_*(\boldsymbol{\delta}^* - \boldsymbol{\delta}) = O_p(K^{-1})$ (as Butar and Lahiri 2003, 74). The proof of (iii) we obtain using

$$\begin{aligned}&\hat{\theta}_{EBLUP}\left(\mathbf{Y}_s, \hat{\beta}\left(\boldsymbol{\delta}^*\right), \boldsymbol{\delta}^*\right) - \hat{\theta}_{EBLUP}(\boldsymbol{\delta}) = \\ &= \left(\boldsymbol{\delta}^* - \boldsymbol{\delta}\right)\frac{\partial c^T(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}}\left(\mathbf{Y}_s - \mathbf{X}_s\hat{\beta}\right) + O_{p^*}\left(K^{-1}\right)\end{aligned}$$

under Model 2, which results directly from Equation (25) in Żądło (2009, 108), which is the direct generalization of Equation (A.2) in Datta and Lahiri (2000, 623) used by Butar and Lahiri (2003, 74). The proof of part (iv) results from the RC, see Butar and Lahiri (2003,75), which implies $\boldsymbol{\delta} - \boldsymbol{\delta} = o_p(1)$, $\Sigma(\boldsymbol{\delta}) = O(K^{-1})$, $\hat{\beta}(\boldsymbol{\delta}) = \beta + o_p(1)$, $\Sigma(\boldsymbol{\delta}) = \Sigma(\boldsymbol{\delta}) + o_p\left(K^{-1}\right)$. Hence, $\frac{\partial c^T(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}} = \frac{\partial c^T(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}} + o_p(1)$ and:

$$\begin{aligned}&\frac{\partial c^T(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}}\left(\mathbf{Y}_s - \mathbf{X}_s\hat{\beta}\right)\left(\mathbf{Y}_s - \mathbf{X}_s\hat{\beta}\right)^T\left(\frac{\partial c^T(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}}\right)^T\Sigma(\boldsymbol{\delta}) = \\ &= \frac{\partial c^T(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}}\left(\mathbf{Y}_s - \mathbf{X}_s\beta\right)\left(\mathbf{Y}_s - \mathbf{X}_s\beta\right)^T\left(\frac{\partial c^T(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}}\right)^T\Sigma(\boldsymbol{\delta}) + \\ &\quad + o_p\left(K^{-1}\right).\end{aligned}$$

Finally, using the expressions for $g_3(\boldsymbol{\delta})$ and $g_4(\boldsymbol{\delta})$ we get the part (iv) of Lemma 2.

Theorem 3. Under Model 1, Model 2 and the RC, we have:

$$E\left(\widehat{MSE}\left(\hat{\theta}_{EBLUP}\right)\right) - MSE\left(\hat{\theta}_{EBLUP}\right) = o\left(K^{-1}\right),$$

where $\widehat{MSE}\left(\hat{\theta}_{EBLUP}\right)$ is given by Equation (19).

Proof. Using Lemma 1 and Lemma 2 we get:

$$
\begin{aligned}
&E(\widehat{MSE}(\hat{\theta}_{EBLUP})) = E\Big[g_1(\hat{\boldsymbol{\delta}}) + g_2(\hat{\delta}) + \\
&-E_*\Big(g_1(\boldsymbol{\delta}^*) + g_2(\boldsymbol{\delta}^*) - (g_1(\hat{\boldsymbol{\delta}}) + g_2(\hat{\boldsymbol{\delta}}))\Big) + \\
&+E_*\big(\hat{\theta}_{EBLUP}(\mathrm{Y}_s, \hat{\beta}(\boldsymbol{\delta}^*), \boldsymbol{\delta}^*) + \\
&-\hat{\theta}_{EBLUP}(\mathrm{Y}_s, \hat{\beta}(\boldsymbol{\delta}), \boldsymbol{\delta})\big)^2\Big] = \\
&= E\Big[g_1(\hat{\boldsymbol{\delta}}) + g_2(\hat{\boldsymbol{\delta}}) - g_1(\hat{\boldsymbol{\delta}}) - \mathbf{B}_{\boldsymbol{\delta}}^T(\boldsymbol{\delta})\frac{\partial g_1(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}} + g_3(\hat{\boldsymbol{\delta}}) + \\
&\quad - g_2(\hat{\boldsymbol{\delta}}) + g_1(\hat{\boldsymbol{\delta}}) + g_2(\hat{\boldsymbol{\delta}}) + g_4(\hat{\boldsymbol{\delta}}) + o_p\big(K^{-1}\big)\Big] \\
&= E\Big[g_1(\hat{\boldsymbol{\delta}}) + g_2(\hat{\boldsymbol{\delta}}) + g_3(\hat{\boldsymbol{\delta}}) + g_4(\hat{\boldsymbol{\delta}}) + \\
&\quad -\mathbf{B}_{\boldsymbol{\delta}}^T(\boldsymbol{\delta})\frac{\partial g_1(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}} + o_p\big(K^{-1}\big)\Big] = \\
&= g_1(\boldsymbol{\delta}) + g_2(\boldsymbol{\delta}) + g_3(\boldsymbol{\delta}) + o\big(K^{-1}\big) \\
&= MSE\big(\hat{\theta}_{EBLUP}\big) + o\big(K^{-1}\big),
\end{aligned}
\tag{24}
$$

where the last equality in Equation (24) for the empirical version of Royall (1976)'s BLUP was proved by Żądło (2009, 110).

5. Real Data Analyses

We will show an application of the proposed method together with other MSE estimators for a real data set. To analyze statistical properties of our method, we will also present Monte Carlo simulation studies, taking the problem of model misspecification into account as well.

In the following analyses, prepared using R (R Core Team 2019), we consider the real population data for $N = 378$ Polish counties called poviats (NUTS 4) for $M = 3$ periods - for the years 2011–2013. Two poviats were excluded from the analysis - the first because of the lack of data, the second (Warsaw) as an outlier. Investments in companies (in hundreds of million PLN) and the number of new companies registered (in hundreds) are the variable of interest and the auxiliary variable, respectively. The data are divided into $D = 28$ domains in the following way. Firstly, the population of poviats (NUTS 4) is divided into 16 voivodships (NUTS 2). Secondly, poviats in each voivodship are divided into two groups according to the type of poviat (city counties and land counties), but only if the sizes of both groups are at least 3. Domain sizes range from 3 to 37 (with the mean: 13.5). The problem of prediction of domain totals in the last period is considered.

In the first period, poviats are divided into two strata. The first stratum consists of poviats from domains that consist only of city counties. Other poviats belong to the second stratum. In the first period, a simple random sample without replacement is drawn from each stratum, (optimal allocation is used), in which the overall sample size is $n = 38$.

Sample sizes in domains range from 0 to 8 (with the mean: 1.36). In 9 domains zero sample sizes are observed. The same elements are in the samples in other periods (the balanced panel sample). This gives the division of the population in each period into the sample and the set of non-sampled elements.

5.1. Application

We mimic the real data analysis for the considered sample data set. We would like to find the model with the best goodness-of-fit measured by AIC and BIC criteria. Of course different measures of goodness-of-fit, including conditional AIC proposed by Vaida and Blanchard (2005) for clustered data, can be used as well. In the example, we do not group the profiles and hence the subscript k is omitted. We study models with one auxiliary variable, with and without random effects, with and without constants that belong to the following two classes. Firstly,

$$Y_{ijd} = \beta_1 x_{idj} + \beta_2 + \upsilon_1 + \upsilon_2 + e_{idj},$$

where $i = 1, 2, ..., N$; $d = 1, 2, ..., D$, $j = 1, 2, ..., M_{id}$; υ_1, υ_2 and e_{idj} are mutually independent, $\upsilon_1 \sim N(0, \sigma_1^2)$, $\upsilon_2 \sim N(0, \sigma_2^2)$ and $e_{idj} \sim N(0, \sigma_e^2)$. Random effects υ_1 and υ_2 (where $\upsilon_1 \neq \upsilon_2$) can be domain specific (υ_d), time specific (υ_t), domain-and-time specific (υ_{dt}) or profile specific (υ_{id}). Secondly,

$$Y_{ijd} = \left(\beta_1 + \upsilon_1\right) x_{idj} + \left(\beta_2 + \upsilon_2\right) e_{idj},$$

where $i = 1, 2, ..., N$; $d = 1, 2, ..., D$, $j = 1, 2, ..., M_{id}$, υ_1, υ_2 and e_{idj} are mutually independent, $\upsilon_1 \sim N\left(0, \sigma_1^2\right), \upsilon_2 \sim N\left(0, \sigma_2^2\right)$ and $e_{idj} \sim N\left(0, \sigma_e^2\right)$. Random effects υ_1 and υ_2 (where $\upsilon_1 \neq \upsilon_2$ or $\upsilon_1 = \upsilon_2$) can be, for example, domain specific (υ_d), time specific (υ_t), domain-and-time specific (υ_{dt}) or profile specific (υ_{id}).

The model with the smallest AIC and BIC criteria is given by

$$Y_{ijd} = (\beta_1 + \upsilon_{id}) x_{idj} + e_{idj}. \tag{25}$$

REML estimates of model parameters are as follows: $\hat{\beta}_1 = 0.4194$, $\hat{\sigma}_\upsilon^2 = 0.0778$ and $\hat{\sigma}_e^2 = 1.5194$. Based on the results of permutation tests we can claim that model parameters are statistically significant (p-values for tests of β_1 and $\hat{\sigma}_\upsilon^2$ are zero). The model belongs to the class of random regression coefficients models considered by Moura and Holt (1999) and Hobza and Morales (2013).

Remark 2. Firstly, in Model (25) the independence of random components (e_{idj}) is assumed, which means that $\mathbf{R}_{rs} = \mathbf{0}$ and hence, predictors (12a) and (14a) are identical and the equalities (16), (17) and (23) are true. Secondly, the model belongs to the class of mixed linear models with independent profile-specific random components, widely discussed by, for example, Verbeke and Molenbergh (2000). In this class of models, nonzero covariances between the variables of interest are observed only within profiles, and hence for the balanced panels samples we have $\mathbf{V}_{rs} = \mathbf{0}$ and (because $\mathbf{G}$ is diagonal)

$\mathbf{Z}_r\mathbf{G}\mathbf{Z}_s^T = \mathbf{0}$. This implies that

$$g_3(\boldsymbol{\delta}) = g_3^s(\boldsymbol{\delta}) = 0,$$

where $g_3(\boldsymbol{\delta})$ and $g_3^s(\boldsymbol{\delta})$ are given by Equations (21) and (20), respectively.

In Table 1 we present values for the following MSE estimators of the empirical version of the predictor (14a):

- the MSE estimator based on the Taylor expansion proposed by Żądło (2009), which is - for the considered model - equivalent to the MSE estimator of the empirical version of the predictor (12a) originally proposed by Datta and Lahiri (2000) (*Taylor)*,
- the MSE estimator based on delete-one-profile jackknife applied for the empirical version of the predictor (14a), based on the idea presented by Jiang et al. (2002) (where delete-one-domain jackknife for empirical version of the predictor (7a) was studied) (*jack)*,
- the MSE estimator based on delete-one-profile weighted-jackknife applied for the empirical version of the predictor (14a), based on the idea presented by Chen and Lahiri (2002, 2003) (where weighted delete-one-domain jackknife for the empirical version of the predictor (7a) was studied) (*w-jack)*,
- the parametric bootstrap MSE estimator studied by González-Manteiga et al. (2007, 2008) (*boot1)*,
- the proposed parametric bootstrap MSE estimator given by (19) (*boot2)*.

Jackknife and weighted jackknife MSE estimators are adapted for the considered predictor by: (i) in formulae presented by Jiang et al. (2002, 1,787) and Chen and Lahiri (2002, 474) replacing g_1^s (.) and g_1^s (.) (see Equations (7e) and (7f) with g_1(.) and g_2(.) (see Equations (14c) and (14d)) and (ii) deleting profiles instead of domains in the case of estimation of $\boldsymbol{\delta}$ (in our model the number of blocks in $\mathbf{V}$ matrix is equal to the number of profiles).

In Table 1 we present results for all observed domain sample sizes. If the same sample size is observed in many domains, we present results only for two of them – with the smallest and the largest MSE estimates. For most of domains, the values of all MSE estimators except *jack* were very similar, which suggests similar stochastic properties (studied in the next section).

Table 1. Values of the MSE estimators for selected domains.

d	n_d	N_d	*Taylor*	*jack*	*w-jack*	*boot1*	*boot2*
13	0	14	36.4	32.2	36.4	36.7	36.5
22	0	3	643.6	612.7	642.0	707.0	643.3
12	1	4	8.1	7.7	8.0	7.8	8.1
6	1	3	798.1	766.9	796.0	824.1	797.8
27	2	4	87.0	83.4	86.8	85.4	87.0
19	2	18	94.5	71.4	94.6	94.2	94.7
11	3	21	85.6	65.4	85.7	89.6	85.8
28	4	21	109.5	87.4	109.7	117.4	109.8
17	6	31	326.1	242.9	326.6	335.2	326.9
8	8	19	145.2	116.1	145.3	142.8	145.5

All of the considered MSE estimators are based on the normal mixed model where the normality of random effects and random components is assumed, which is equivalent (logical biconditional) to normality $\mathbf{Y}$ (see Remark 3).

Remark 3. Assuming (3) (including independence of $\mathbf{v}$ and $\mathbf{e}$ and full column rank of $\mathbf{Z}$), $\mathbf{v}$ and $\mathbf{e}$ have multivariate normal distributions if and only if $\mathbf{Y}$ has the multivariate normal distribution. The above statement is true because both implications are true. The implication, that if $\mathbf{v}$ and $\mathbf{e}$ in Equation (3) have multivariate normal distributions, then $\mathbf{Y}$ has the multivariate normal distribution, is a standard result (see e.g., Muirhead (2005), 6 theorem 1.2.6 and 14 theorem 1.2.14). The implication, that if $\mathbf{Y}$ in (3) has the multivariate normal distribution, then $\mathbf{v}$ and $\mathbf{e}$ have multivariate normal distribution, results from the following two statements. Firstly, if $\mathbf{Y}$ in (3) has the multivariate normal distribution, then two random vectors of the same size $\mathbf{Zv}$ and $\mathbf{e}$ have multivariate normal distributions (see e.g., Muirhead (2005), 14 theorem 1.2.13). Secondly, if $\mathbf{Zv}$ has multivariate normal distribution and $\mathbf{Z}$ is of full rank, then $\mathbf{v}$ has the multivariate normal distribution. It is true, because there exists a linear transformation of $\mathbf{Zv}$ to $\mathbf{v}$ (e.g., $((\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T)(\mathbf{Zv}) = \mathbf{v}$) and then (e.g., Muirhead (2005), 6 theorem 1.2.6) normality of $\mathbf{Zv}$ implies normality of $\mathbf{v}$.

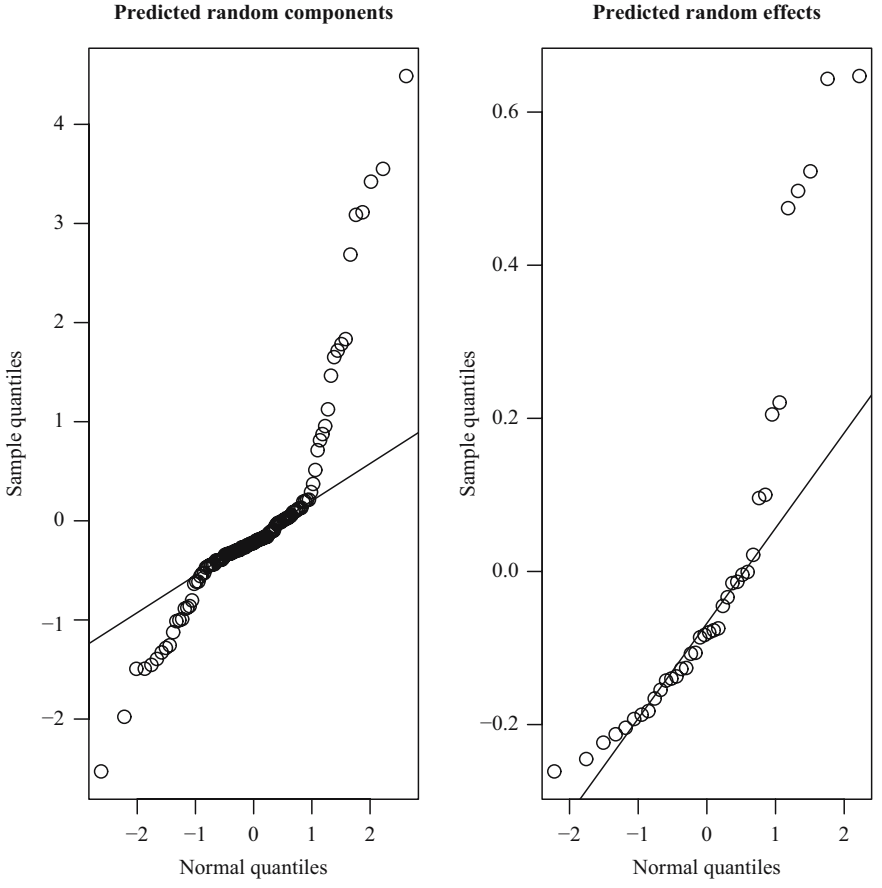


Fig. 1. Q-Q plots of predicted random components and random effects.

The crucial point of the proofs of asymptotic unbiasedness of MSE estimators, where normality of $\mathbf{Y}$ (normality of random effects and random components) is required, is the approximate decomposition of the MSE of the EBLUP into: the MSE of the BLUP and the additional component usually denoted by $g_3(.)$ (see Equation (20) for our case). The proof of the decomposition for the empirical version of the predictor studied by Henderson (1950) is considered by, for example, Kackar and Harville (1984, 855) and Robinson (1991, 19) and for the empirical version of the predictor studied by Royall (1976) and by Harville and Jeske (1992) in Section 2. Then, based on the decomposition (under normality of $\mathbf{Y}$), $g_3(.)$ component is estimated using different methods and, finally, the MSE estimator is obtained. The assumption of normality of $\mathbf{Y}$ is used in the case of the MSE estimator based on Taylor's expansion by Datta and Lahiri (2000, 623) and Żądło (2009), for the weighted jackknife MSE estimator by Chen and Lahiri (2003, 908) and for the parametric bootstrap MSE estimators by Butar and Lahiri (2003, 66) and in our proposal (see Remark 1). In case of the jackknife MSE estimator the proof is different, but normality of $\mathbf{Y}$ is also required, as shown by Jiang et al. (2002, 1,803). For our data, the assumption of normality of $\mathbf{Y}$ (or equivalently the normality of $\mathbf{v}$ and $\mathbf{e}$ – see Remark 3) is not met – using the Shapiro test we checked normality using residuals $\mathbf{Y}_s - \mathbf{X}_s\hat{\beta}$ after the Cholesky transformation (p-value equals 0). Q-Q plots of predicted random effects $\hat{\mathrm{v}} = \tilde{\mathrm{v}}(\hat{\boldsymbol{\delta}})$ (where $\tilde{\mathrm{v}}(\boldsymbol{\delta})$ is given by Equation (7c)) and predicted random components (conditional residuals) $\hat{\mathrm{e}}_s = \mathbf{Y}_s - \mathbf{X}_s\hat{\beta} - \mathbf{Z}_s\hat{\mathbf{v}}$ are presented in Figure 1. Hence, in the next section we will study the biases and MSEs of the considered MSE estimators under normality and different non-normal cases. We are interested in comparing the behavior of our estimator with its competitors under different distributions.

5.2. *Simulation Study*

In the model-based simulation study prepared in R (R Core Team 2019), we analyze the same MSE estimators, the same data, the same model and the same division into the sampled and non-sampled sets as in the previous section. The number of iteration equals $L = 5{,}000$.

Bootstrap MSE estimators are computed based on 200 replications.

The values of variable of interest are generated based on Model (25) (with parameters computed based on the whole population data), where random effects and random components are generated using the following distributions:

- normal,
- scaled t-Student with 3 degrees of freedom,
- shifted exponential,
- shifted log-normal (where the third standardized moment equals 3),
- shifted gamma (where the third standardized moment equals 4),
- shifted Pareto (where the third standardized moment equals 5).

Distributions with positive asymmetry are usually used in Monte Carlo simulation studies in economic applications (e.g., Białek 2014). The true MSEs are computed based on the following formula: $MSE = L^{-1}\sum_{l=1}^{L}\left(\hat{\theta}_l - \theta_l\right)^2$, where the number of Monte Carlo iterations equals $L = 5000$, $\hat{\theta}_l$ and θ_l are computed for lth iteration as Equations (14a) and (5a), respectively.

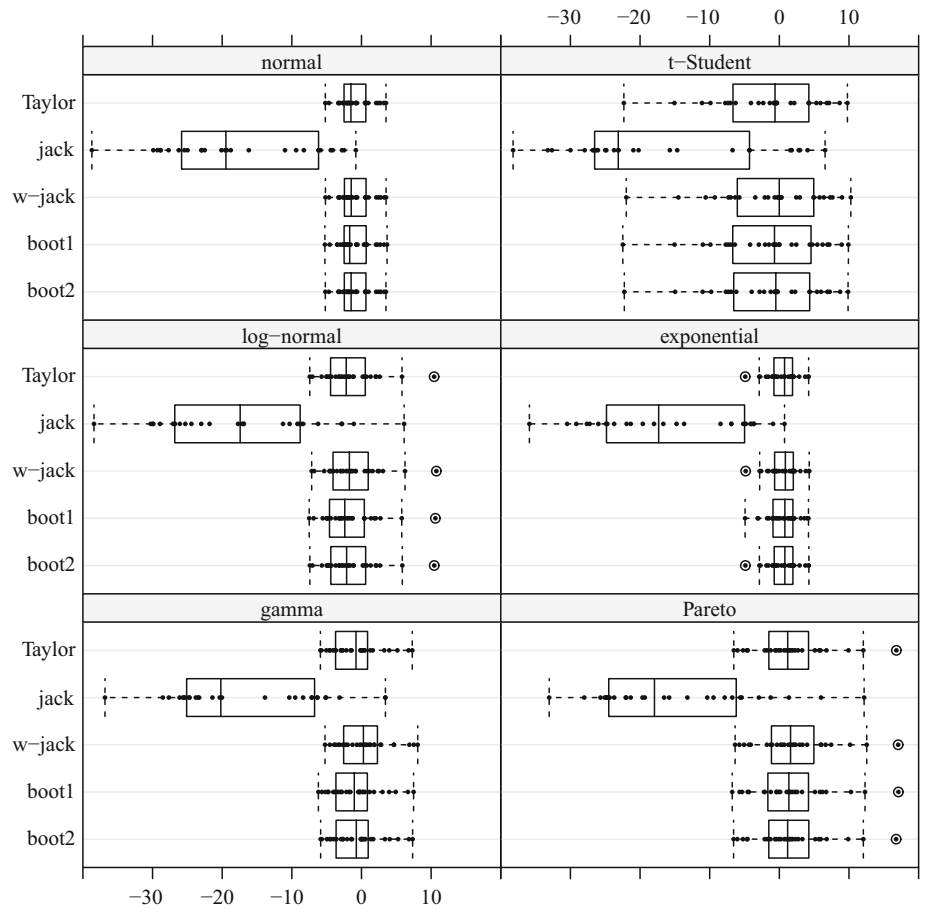


Fig. 2. Relative biases of MSE estimators for D = 28 domains (in percent).

In Figure 2 we present relative biases of the MSE estimators computed as $rB\left(\widehat{MSE}\right) = MSE^{-1}L^{-1}\sum_{l=1}^{L}\left(\widehat{MSE}_l - MSE\right)$, where $\widehat{MSE}_l$ is the value of the MSE estimator computed for the lth iteration. If the normality assumption is met, the relative biases of the estimators, except *jack*, are between approximately -5% and 4%. For other distributions, absolute biases are larger, but - in our opinion - acceptable for most of the domains. Moreover, median relative biases for all of the MSE estimators except *jack* are very close to zero.

Even though for non-normal distributions the increase of the absolute biases is in our opinion and this acceptable, we observe a large increase of relative RMSEs (see Figure 3) computed as $rRMSE\left(\widehat{MSE}\right) = MSE^{-1}\left(L^{-1}\sum_{l=1}^{L}(\widehat{MSE}_l - MSE)^2\right)^{0.5}$. Values of relative RMSEs for normal distribution are between approximately 18% and 29% (except *jack)*, but for other distributions they even exceed 100% for some cases. It is interesting to note that although large absolute biases are observed for the *jack* MSE estimator (see Figure 2) it is the MSE estimator with the smallest RMSEs for non-normal distributions (see Figure 3).

Summing up the results of the Monte Carlo simulation study presented in the previous two paragraphs, we can state that all of the considered MSEs are robust on the lack of

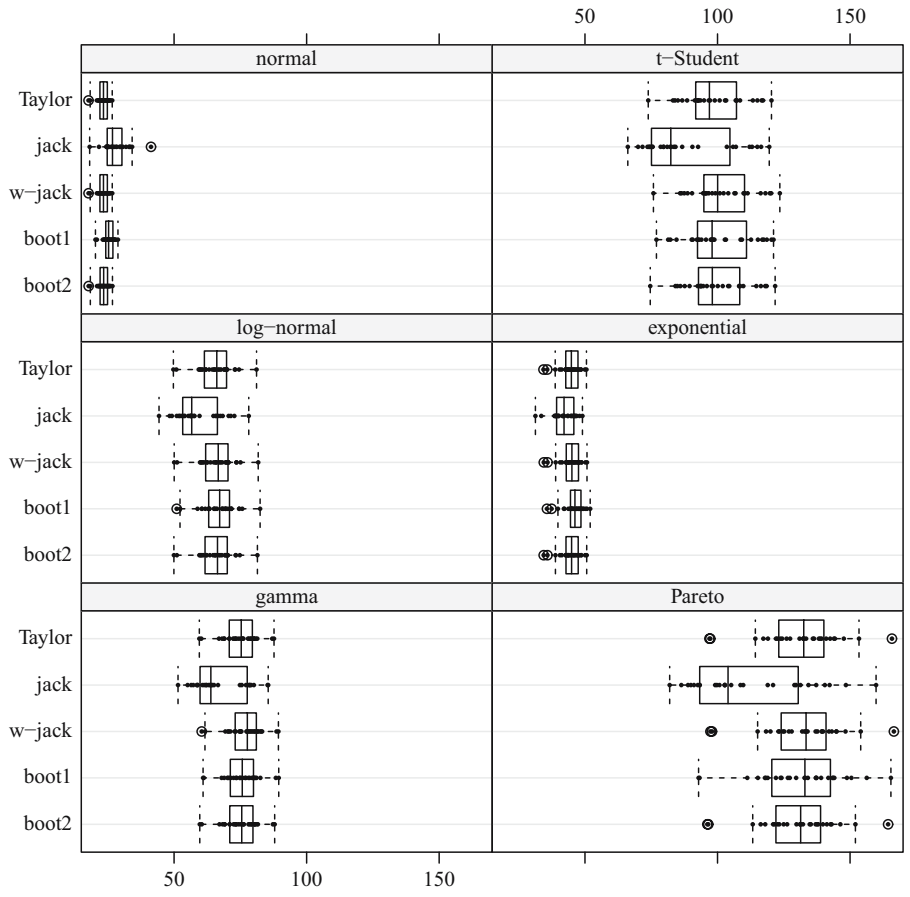


Fig. 3. Relative RMSEs of MSE estimators for $D = 28$ domains (in percent).

normality taking into account their biases, but not if their MSEs are considered. We observe a large decrease of their accuracies both for the heavy-tailed scaled t-Student distribution and for positively skewed distributions, where for the stronger asymmetry a larger increase of the MSE is observed. Interesting results are observed for the jackknife MSE estimator for which slightly smaller MSEs are received in non-normal cases. Furthermore, the proposed parametric bootstrap MSE estimator, jackknife and weighted jackknife MSE estimators (that we adapted for the considered predictor) have very similar properties compared with the MSE estimator-based on the Taylor expansion. However, their advantage over the Taylor expansion-based MSE estimator is that they do not require derivation of the $g_3(\boldsymbol{\delta})$ component of the MSE, which can be problematic for complex superpopulation models.

6. Conclusion

In the article, we show that the empirical version of the best linear unbiased predictor proposed by Royall (1976) is a generalization of the empirical version of the predictor

studied by Henderson (1950), together with the condition when they are equivalent. We generalize the parametric bootstrap MSE estimator proposed by Butar and Lahiri (2003) and prove its asymptotic unbiasedness. The proof of asymptotic unbiasedness requires normality of random effects and random components, which is a limitation of the method but is also a typical assumption for other MSE estimators. This is the reason why, in the application and in the simulation study, the properties of the proposed MSE estimator are compared with other MSE estimators, showing that they have very similar properties both under normality and non-normal cases. However, our method has the advantage that it does not require derivation of the $g_3(.)$ component of the MSE, as the MSE estimator is based on Taylor's expansion, which can be crucial for models with a complex covariance structure.

7. Appendix

We assume the following regularity conditions, which will be referred to as (RC). The following regularity conditions (a), (b), (e), (f), (g) are direct generalizations of the respective regularity conditions proposed by Butar and Lahiri (2003) for the proposed longitudinal model (i.e., the definition sizes of the matrices cover the case of longitudinal data and the assumption that population and subpopulations can change in time). We replace regularity conditions (c) and (d) proposed by Butar and Lahiri (2003) with the following regularity conditions (c) and (d) to cover both the more general predictor (the empirical version of the predictor proposed by Royall 1976) and the proposed model. The regularity conditions (RC) are:

(a) The elements of matrices $\mathbf{X}_k$ and $\mathbf{Z}_k$ are uniformly bounded such that $\left\{\mathbf{X}_s^T\mathbf{V}_{ss}^{-1}\mathbf{X}_s\right\} = [O(K)]_{p\times p}$,

(b) $\sup_{1\le d\le D} n_k < \infty$ and $\sup_{1\le d\le D} r_k < \infty$,

(c) $\mathbf{X}_r^T\gamma_r - \mathbf{X}_s^T\mathbf{V}_{ss}^{-1}\mathbf{V}_{sr}\gamma_r = [O(1)]_{p\times 1}$,

(d) $\frac{\partial}{\partial\boldsymbol{\delta}_c}\mathbf{X}_s^T\mathbf{V}_{ss}^{-1}\mathrm{V}_{sr}\gamma_r = [O(1)]_{p\times 1}$ where $c = 1, 2, \ldots, q$,

(e) $\mathbf{R}_{sk}(\boldsymbol{\delta}) = \sum_{c=0}^{q}\boldsymbol{\delta}_j\mathbf{C}_{kc}\mathbf{C}_{kc}^T$ and $\mathbf{G}_k(\boldsymbol{\delta}) = \sum_{c=0}^{q}\boldsymbol{\delta}_j\mathbf{F}_{kc}\mathbf{F}_{kc}^T$ where $\boldsymbol{\delta}_0 = 1$, $\mathbf{C}_{kc}$ and $\mathbf{F}_{kc}$ $(k = 1, 2,..., K;\ c = 0, 1,..., q)$ are known matrices of the order $n_k \times r_k$ and $r_k \times r_k$, respectively, and the elements are uniformly bounded known constants such that $\mathbf{R}_{sk}(\boldsymbol{\delta})$ and $\mathbf{G}_k(\delta)$ $(k = 1, 2,..., K)$ are all positive definite matrices. In special cases, some of $\mathbf{C}_{kc}$ and $\mathbf{F}_{kc}$ may be null matrices.

(f) $\hat{\boldsymbol{\delta}}$ is an estimator of $\boldsymbol{\delta}$ that satisfies (i) $\hat{\boldsymbol{\delta}} - \boldsymbol{\delta} = O_p(K^{-1/2})$, (ii) $\hat{\boldsymbol{\delta}} - \hat{\boldsymbol{\delta}}_{ML} = O_p(K^{-1})$ (where $\hat{\boldsymbol{\delta}}_{ML}$ is the maximum-likelihood estimator of $\boldsymbol{\delta}$), (iii) $\hat{\boldsymbol{\delta}}(\mathbf{Y}_s) = \hat{\boldsymbol{\delta}}(-\mathbf{Y}_s)$, (iv) $\hat{\boldsymbol{\delta}}(\mathbf{Y}_s + \mathbf{X}_s\mathrm{b}) = \hat{\boldsymbol{\delta}}(\mathbf{Y}_s)$, for any $\mathrm{b} \in R^p$ and for all $\mathbf{Y}_s$. Assume that $E(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}) = \mathrm{B}_{\boldsymbol{\delta}}(\boldsymbol{\delta}) + o(K^{-1})$, which means that the approximate (to the order $o(K^{-1})$) formula of the bias of $\boldsymbol{\delta}$ is known.

(g) $E(\hat{\beta}(\boldsymbol{\delta}) - \beta)(\boldsymbol{\delta} - \boldsymbol{\delta})^T = o(K^{-1})$.

8. References

Białek, J. 2014. "Simulation study of an original price index formula." *Communications in Statistics - Simulation and Computation* 43: 285–297. DOI: http://doi.org/10.1080/03610918.2012.700367.

Butar, F.B. and P. Lahiri. 2003. "On measures of uncertainty of empirical Bayes small-area estimators." *Journal of Statistical Planning and Inference* 112: 63–76. DOI: http://doi.org/10.1016/S0378-3758(02)00323-3.

Chen, S. and P. Lahiri, 2002. "A weighted jackknife MSPE estimator in small-area estimation." In Proceedings of the Section on Survey Research Methods: American Statistical Association, May 14–19, 2002: 473–477. Florida, VAL: American Statistical Association. Available at: http://www.asasrms.org/Proceedings/y2002/Files/JSM2002-001127.pdf (accessed January 2020).

Chen, S. and P. Lahiri. 2003. "A comparison of different MSPE estimators of EBLUP for the Fay-Herriot model." In Proceedings of the Section on Survey Research Methods: American Statistical Association, May 15–18, 2003: 905–911. Nashville, VAL: American Statistical Association. Available at: http://www.asasrms.org/Proceedings/y2003/Files/JSM2003-000585.pdf (accessed January 2020).

D'Aló, M., S. Falorsi, and F. Solari. 2017. "Space-Time Unit-Level EBLUP for Large Data Sets." *Journal of Official Statistics* 33: 61–77. DOI: http://doi.org/10.1515/jos-2017-0004.

Das, K., J. Jiang, and J.N.K. Rao. 2004. "Mean squared error of empirical predictor." *The Annals of Statistics* 32: 818–840. DOI: http://doi.org/10.1214/009053604000000201.

Datta, G.S. and P. Lahiri. 2000. "A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems." *Statistica Sinica* 10: 613–627. Available at: http://www3.stat.sinica.edu.tw/statistica/oldpdf/A10n214.pdf (accessed January 2020).

Fabrizi, E., M.R. Ferrante, and S. Pacei. 2007. "Small area estimation of average household income based on unit level models for panel data." *Survey Methodology* 33: 187–198. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2007002/article/10496-eng.pdf?st=eifh–iS (accessed January 2020).

González-Manteiga, W., M.J. Lombardía, I. Molina, D. Morales, and L. Santamaría. 2007. "Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model." *Computational Statistics and Data Analysis* 51: 2720–2733. DOI: http://doi.org/10.1016/j.csda.2006.01.012.

González-Manteiga, W., M.J. Lombardía, I. Molina, D. Morales, and L. Santamaría. 2008. "Bootstrap mean squared error of small-area EBLUP." *Journal of Statistical Computation and Simulation* 78: 443–462. DOI: http://doi.org/10.1080/00949650601141811.

Harville, D.A. and D.R. Jeske. 1992. "Mean square error of estimation or prediction under general linear model." *Journal of Statistical Computation and Simulation* 87: 724–731. DOI: http://doi.org/10.1080/01621459.1992.10475274.

Henderson, C.R. 1950. "Estimation of genetic parameters (Abstract)." *Annals of Mathematical Statistics* 21: 309–310.

Hobza, T. and D. Morales. 2013. "Small area estimation under random regression coefficient models." *Journal of Statistical Computation and Simulation* 83: 2160–2177. DOI: http://doi.org/10.1080/00949655.2012.684094.

Hobza, T., D. Morales, and L. Santamaría. 2018. "Small area estimation of poverty proportions under unit-level temporal binomial-logit mixed models." *TEST* 27: 270–294. DOI: http://doi.org/10.1007/s11749-017-0545-3.

Jiang, J. 1996. "REML Estimation: Asymptotic Behavior and Related Topics." *The Annals of Statistics* 24: 255–286. DOI: http://doi.org/10.1214/aos/1033066209.

Jiang, J. and P. Lahiri. 2006. "Mixed model prediction and small area estimation." *Test* 15: 1–96. DOI: http://doi.org/10.1007/BF02595419.

Jiang, J., P. Lahiri, and S.-M. Wan. 2002. "Unified jackknife theory for empirical best prediction with M-estimation." *The Annals of Statistics* 30: 1782–1810. DOI: http://doi.org/10.1214/aos/1043351257.

Kackar, R.N. and D.A. Harville. 1981. "Unbiasedness of two-stage estimation and prediction procedures for mixed linear models." *Communications in Statistics, Ser. A* 10: 1249–1261. DOI: http://doi.org/10.1080/03610928108828108.

Kackar, R.N. and D.A. Harville. 1984. "Approximations for standard errors of estimators of fixed and random effects in mixed linear models." *Journal of the American Statistical Association* 79: 853–862. DOI: http://doi.org/10.1080/01621459.1984.10477102.

Marhuenda, Y., I. Molina, and D. Morales. 2013. "Small area estimation with spatio-temporal Fay-Herriot models." *Computational Statistics & Data Analysis* 58: 308–325. DOI: http://doi.org/10.1016/j.csda.2012.09.002.

Molina, I., D. Morales, M. Pratesi, and N. Tzavidis. eds. 2010. "Final small area estimation developments and simulation results." *SAMPLE project: Small Area Methods for Poverty and Living Conditions Estimates*. http://www.sample-project.eu/images/stories/docs/samplewp2d12%2616_saefinal.pdf (accessed January 2020).

Moura, F.A.S. and D. Holt. 1999. "Small area estimation using multilevel models." *Survey Methodology* 25: 73–80.

Morales, D. and L. Santamaría. 2019. "Small area estimation under unit-level temporal linear mixed models." *Journal of Statistical Computation and Simulation.* 89: 1592–1620. DOI: http://doi.org/10.1080/00949655.2019.1590578.

Muirhead, R.J. 2005. *Aspects of Multivariate Statistical Theory*. New Jersey: John Wiley & Sons.

Nissinen, K. 2009. *Small area estimation with linear mixed models for unit-level panel and rotating panel data*. Dissertation, University of Jyväskylä.

Pagliarella, M.C. and R. Salvatore. 2016. "Unit Level Spatio-temporal Models." In *Analysis of Poverty Data by Small Area Estimation*, edited by M. Pratesi, 227–243. Chichester: Wiley.

Prasad, N.G.N. and J.N.K. Rao. 1990. "The estimation of mean the mean squared error of small-area-estimators." *Journal of the American Statistical Association* 85: 163–171. DOI: http://doi.org/10.1080/01621459.1990.10475320.

R Core Team. 2019. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. Available at: https://www.R-project.org/ (accessed January 2020).

Rao, J.N.K. and I. Molina. 2015. *Small area estimation. Second edition*. New Jersey: Wiley.

Rao, J.N.K. and M. Yu. 1994. "Small area estimation by combining time-series and cross-sectional data." *Canadian Journal of Statistics* 22: 511–528. DOI: http://doi.org/10.2307/3315407.

Robinson, G.K. 1991. "That BLUP is a good thing: the estimation of random effects." *Statistical Science* 6: 15–51. DOI: http://doi.org/10.1214/ss/1177011926.

Royall, R.M. 1976. "The linear least squares prediction approach to two-stage sampling." *Journal of the American Statistical Association* 71: 657–473. DOI: http://doi.org/10.1080/01621459.1976.10481542.

Saei, A. and R. Chambers. 2003. *Small Area Estimation Under Linear and Generalized Linear Mixed Models with Time and Area Effects*. S3RI, Methodology Working Paper M03/15r, Southampton: University of Southampton. Available at: https://eprints.soton.ac.uk/8165/1/8165-01.pdf (accessed January 2020).

Schmid, T. and R. Münnich. 2014. "Spatial robust small area estimation." *Statistical Papers* 55: 653–670. DOI: http://doi.org/10.1007/s00362-013-0517-y.

Stukel, D.M. and J.N.K. Rao. 1999. "On small-area estimation under two-fold nested error regression models." *Journal of Statistical Planning and Inference* 78: 131–147. DOI: http://doi.org/10.1016/S0378-3758(98)00211-0.

Vaida, F. and S. Blanchard. 2005. "Conditional Akaike Information for Mixed-Effects Models." *Biometrika* 92: 351–370. DOI: http://doi.org/10.1093/biomet/92.2.351.

Valliant, R., A.H. Dorfman, and R.M. Royall. 2000. *Finite population sampling and inference. A prediction approach*. New York: John Wiley & Sons.

Verbeke, G. and G. Molenberghs. 2000. *Linear mixed models for longitudinal data*. New York: Springer-Verlag.

Żądło, T. 2004. "On unbiasedness of some EBLU predictor." In *Proceedings in Computational Statistics 2004*, Physica-Verlag, Heidelberg-New York, August 23–27, 2004. 2019–2026. Prague, VAL: International Association for Statistical Computing. Available at: https://link.springer.com/book/10.1007/978-3-7908-2656-2 (accessed January 2020).

Żądło, T. 2009. "On MSE of EBLUP." *Statistical Papers* 50: 101–118. DOI: http://doi.org/10.1007/s00362-007-0066-3.

Żądło, T. 2014. "On the Prediction of the Subpopulation Total Based on Spatially Correlated Longitudinal Data." *Mathematical Population Studies* 21: 30–44. DOI: http://doi.org/10.1080/08898480.2013.836387.

Żądło, T. 2015a. "On longitudinal moving average model for prediction of subpopulation total." *Statistical Papers* 56: 749–771. DOI: http://doi.org/10.1007/s00362-014-0607-5.

Żądło, T. 2015b. "On prediction for correlated domains in longitudinal surveys." *Communications in Statistics - Theory and Methods* 44: 683–697. DOI: http://doi.org/10.1080/03610926.2013.857867.

Received April 2019
Revised October 2019
Accepted January 2020

sciendo

Journal of Official Statistics, Vol. 36, No. 2, 2020, pp. 459–461, http://dx.doi.org/10.2478/JOS-2020-0023

Book Review

Peter Struijs[1]

Boris Lorenc, Paul A. Smith, Mojca Bavdaž, Gustav Haraldsen, Desislava Nedyalkova, Li-Chun Zhang and Thomas Zimmermann, eds. *The Unit Problem and Other Current Topics in Business Survey Methodology.* 2018 Newcastle upon Tyne: Cambridge Scholars Publishing, ISBN 978-1-5275-1661-8, 288 pages.

In business survey methodology, increasing attention is given to the "unit problem". The unit problem is the set of issues associated with the application of the statistical concept of the unit in business statistics. Various types of units may be defined and applied by statisticians, administrative registers and businesses themselves, which greatly complicates the production of statistics. Issues include the delineation of units, for instance if an enterprise is composed of several legal units. They include relating collection and observation units to the target units of business statistics, for instance if local units are used for collecting data on businesses. They also include identifying the units to be listed in business registers, especially if different types of units are used in business statistics, as is the case in the European Statistical System (ESS). Next to the enterprise, which is defined as an independent business actor in the economy, the ESS makes use of, for instance, the kind-of-activity unit, which is roughly defined as part of an enterprise carrying out a single industrial activity and administered as an entity in its own right. There are associated issues of sampling methodology, of coordination and integration of business statistics, of costs, efficiency and data availability, and more. Furthermore, all issues can be seen from the perspective of the errors they may cause in the statistics produced.

Given the high relevance of the unit problem, the book *The Unit Problem and Other Current Topics in Business Survey Methodology* by Boris Lorenc *et al*. (eds) is a very welcome contribution to understanding such issues. However, the book, which comprises an edited and enriched selection of papers presented at the 2017 European Establishment Statistics Workshop, also covers other business survey topics, related or less related. Examples include sampling coordination, managing response burden, questionnaire design, using new data sources for price statistics, and data visualization. The book will not only benefit survey methodologists and producers of business statistics in general, but also the users of such statistics, widening their understanding of business statistics and their intricacies.

Roughly half the book is concerned with the unit problem. The unit problem itself is explained in a chapter by Smith, Lorenc and van Delden, in which they consider unit errors in the context of the five dimensions of output quality as applied in the ESS.

[1] Statistics Netherlands, CBS weg, Heerlen, 6401 CZ. Netherlands. Email: p.struijs@cbs.nl

These are user relevance, accuracy and reliability, timeliness and punctuality, accessibility and clarity, and coherence and comparability. Such a quality framework is helpful in dealing with the unit problem in a structured way. Sturm discusses the unit problem in the context of business registers in Germany, where profiling (i.e., unit delineation) is being implemented. Important issues are the relationship between the enterprise and the kind-of-activity unit as defined and applied in the ESS, and the definition of the enterprise itself. The discussion in the ESS on these matters has not yet been settled.

The book includes several French contributions touching on various aspects of the unit problem. Haag explains how the French business register (actually a network of registers) makes a distinction between enterprises and legal units. The latter used to be the unit applied in business statistics, but since 2010 a change to using the enterprise has been carried through, as that unit has more economic relevance, especially if business structures are complex. Haag quantifies the differences between enterprise and legal unit populations, thereby illustrating the importance of making the distinction. Two-stage cluster sampling is applied to deal with the difference between the collection unit and the statistical unit, as explained by Gros and Le Gleut. The design optimizes the statistics produced at the level of enterprises under a constraint on the number of legal units surveyed. A second contribution of Gros and Le Gleut explains how samples for different surveys are coordinated, positively and negatively, using permanent random numbers. The transition from legal units to enterprises as the unit of business statistics has consequences for the treatment of influential values, that is, winsorization. This is the topic of a contribution by Fizzala.

Four more chapters are clearly linked to the unit problem. Van Delden looks at issues arising when statistics are based on integration of various data sources. These may have different unit types. By using a linkage and data integration framework, he identifies the issues that occur in twelve Dutch case studies. Lammers discusses ways to improve the efficiency of profiling in the Dutch business register. The analyses of the process are based on process mining techniques, which involve the generation of mostly quantitative metadata from the applications used. The unit problem also has a regional dimension, which Ichim discusses for the Italian case of producing business indicators using territorial domains. This requires a strategy for linking enterprises to the local unit level. An analysis of the response process for Norwegian business statistics is provided by Haraldsen, in which the risks to data quality are central. These risks are related to the complexity of the unit structure, which also largely determines the response burden.

Concerning the other topics of the book, two chapters are about sampling. Zimmermann, Schmiedel and Lorentz show how the German federal statistical institute responded to a court decision requiring a much more even spread of the response burden for the services sector. A new sampling design was developed, but take-all strata could not entirely be avoided. The Dutch methodology of sampling coordination across all business surveys and panels is described in a contribution by Smeets and Boonstra. Use is made of permanent random numbers and response burden values. Two more chapters concern data collection. For the United Kingdom, Steward, Sidney and Timm looked at paradata generated during the completion of questionnaires in order to improve them. Both quantitative paradata was used, such as the number of error messages, and qualitative paradata, such as call records. Distinguishing between different respondent types is key to

their analysis. A Swedish experiment with validation embedded in an electronic data collection tool is the subject of a chapter by Lorenc, Norberg and Ohlsson. The experiment sheds not only light on data quality aspects, but also on the relationship between data validation, response burden and costs.

Another topic of the book is the change in the use of data sources for price statistics. This topic is introduced by Zhang, who mentions the increased use of web scraping and scanner data, in addition to the more traditional data collection through enterprise reporting. As a consequence of the change, issues arise concerning the index formula to be used. A Canadian investigation of the possible use of scanner data and the challenges posed is described by Deshaies-Moreault, Harper and Yung. The current CPI approach is compared to one in which scanner data plays a more prominent role. The Slovenian approach to price surveys is explained in a contribution by Razinger, including new developments concerning data collection methods. The last topic of the book, which is also relevant outside the realm of business statistics, is data visualization. Vila, Cervera-Ferri, Camões, Bolko and Bavdaž make a case for its relevance to statistics, and argue that good visualization requires a worked-out methodology in which the cognitive processes underlying data interpretation play a central role. This can be studied empirically, for instance by studying alternative presentations and using eye tracking and gamification.

The Unit Problem and Other Current Topics in Business Survey Methodology is a rich book that is accessible and worth reading by all interested in business survey methodology, including the users of business statistics. Moreover, its special focus on the unit problem and its consequences for business statistics makes it unique.