The quality of estimates in time-series inference problems based on non-probability samples: Enhancing selection bias estimation through historical information

Methodology and Statistics – Faculty of Social Sciences, Utrecht University

Master Thesis – Midterm Research Report

Author: Marco Ramljak (6899587)

Supervisors: Prof. Dr. Ton de Waal (CBS), Dr. Arnout van Delden (CBS)

Date: January 11, 2021

Word Count: 4.897 (excluding references)

Table of contents

Introduction

Inferences based on non-probability samples are increasingly found in news outlets, academic research and commercial research, even though the quality of these estimates (e.g., the presence of selection bias) is still under heavy scrutiny (see e.g., Valliant 2020; Cornesse et al. 2020; Callegaro et al. 2014). The current incentives for using non-probability samples rather than the usual probability samples (design-based samples), are manifold, e.g., they can be cheaper, more flexible and include more units. One might think that including more units (i.e., increasing the sample size) makes population inference easier and more robust – potentially solving the problem of increasing non-response rates in traditional probability samples (De Broe et al. 2020; Kalton 2019; Brick and Williams 2013).

However, current research suggests that non-probability samples suffer from the same problem only from a different point of view: Selectivity is introduced not because certain sampled observational units are missing, rather, selectivity is introduced because only certain observational units are available. The term "certain" emphasizes that we expect the existence of selectivity with respect to a target variable purely because of differing and unknown inclusion probabilities between the observational units. This means that non-probability samples face the same kind of challenge in terms of selectivity as probability samples. Therefore, further research needs to be invested in estimating the degree of selection bias (selectivity) – particularly in inference problems based on non-probability samples, as these are even more and more used in fields like Official Statistics. In the production of Official Statistics unbiased estimates are of particular importance as they form the basis for policy making.

To simplify bias estimation in non-probability samples, different assumptions and auxiliary information can be used. Current measures utilize, for example, auxiliary variables that are known for the whole population to evaluate the representativity of the incompletely measured target variable in the sample. Simulation studies have shown that this is a viable approach in measuring the bias (Boonstra et al. 2019), however, we expect to increase performance when implementing further proxy data, such as historical information. For example, numerous Official Statistics' productions are

conducted where estimations are repeated monthly or yearly by using time-series data (in the following time-series inference problems). Given the stability of these periodic estimations one can try to implement this historical information (previous periods) into the estimation of the selection bias for the subsequent period. In this thesis project we are therefore particularly interested in the estimation of selection bias in time-series inference problems that are based on non-probability samples and focus on finding out which estimation measure performs best.

The paper at hand constitutes a mid-term research report for a thesis project due in May 2021. For the overall thesis project we will, first, introduce two recent frameworks that focus on selection bias in non-probability research – the Standardized Measure for Unadjusted Bias by Little, West, Boonstra and Hu (2020) and the data defect index by Meng (2018). Here, we will incorporate historical information on selection bias into the introduced frameworks, in order to adapt them to time-series inference problems. Second, we will compare and showcase their performance of these potentially "enhanced" measures in a real case example based on an administrative data source. We will use short term business statistics for measuring the average turnover of Dutch businesses using quarter yearly periods. This data has the advantage that observational units are gradually added over time within a certain quarter, giving us the opportunity in facing various degrees of selectivity within a quarter-specific inference problem. At the end of each quarter, ground truth values for the level of selectivity can be derived, which provide the basis for performance evaluation of both measures as well as further historical information to implement into the bias estimation of the next quarter. These properties enable a framework in which the features of non-probability samples are very well mimicked, making it an excellent case study for further research in the field. Furthermore, the case study will highlight the potential flexibility of the tested selection bias estimation measures in incorporating additional information.

The research report at hand resembles the first part of the thesis project – introducing and comparing the two frameworks, partially formalizing the corresponding enhanced measures as well as explaining the plan for the case study. The report is structured as follows. In Section 2, we will introduce the main notation and necessary assumptions as well as important explain the concept of selectivity.

Section 3 will focus on the two frameworks and estimators by Meng (2018) and Little et al. (2020), respectively. Furthermore, Section 3 will present the current status of the formalized versions of the enhanced measures. Finally, Section 4 will discuss the mentioned case study and its parameters.

Defining Selectivity

We are interested in primarily estimating selection bias (also referred to as selectivity) when inferring the population mean of a certain target variable based on a non-probability sample in the context of time-serial data.[1]

Suppose a non-probability dataset $\boldsymbol{D} = \{y_i, z_i, i = 1, \dots, n\}$, drawn from a finite population $U = 1, \dots, N$, with i as the observational unit of analysis and a sample size of n. Z describes a vector of auxiliary variables that are available for the whole population. Y is the continuous target variable of interest, only available for selected observations. Let the selection be described by the dichotomous vector S, where $S = 1$ if the observational unit i in the sample has a measured value for Y, i.e., it was selected in $\boldsymbol{D}$, and $S = 0$ if the value is missing/not available for Y, i.e., it was not selected in $\boldsymbol{D}$. A superscript in brackets will denote group membership of i concerning selection, for example, $\bar{y}^{(1)}$ will describe the mean of Y considering only the selected units and $\bar{y}^{(0)}$ the mean of Y considering only the non-selected units. $\bar{Y}$ will describe the mean of the whole population for Y. This notation logic also applies for variables of Z, even though it is assumed that complete information for auxiliary variables on all units is available. Furthermore, we assume the following:

- The non-probability sample $\boldsymbol{D}$ contains no measurement or linkage errors.
- The target population U is fully known.
- Auxiliary variables are available for all units in U and correlate either with the selection variable S or with the target variable Y.
- The primary estimand of interest is the average outcome of the target variable Y: $E(Y_i) = \bar{Y}$.

The fundamental problem with using non-probability samples for population inference is that sample statistics, e.g., the sample mean, are <u>not</u> an unbiased estimator of the population mean. In contrast to design-based probability samples, with non-probability samples we are not in control which

[1] The notation in this report follows the style of Little et al.(2020).

units are actually sampled, leading to the fact that the sample might not be representative for the population. In terms of Rubin's (1976) missing data framework, non-probability sampling might lead to *non-ignorable* selection bias. To overcome this problem current research is developing measures and methods which try to quantify and ultimately correct for selection bias.

Little et al. (2020) introduce their measure for selection bias with a re-specification of Rubin's framework for missing data (1976) – they shifted their focus from the missing observations to the selected cases, i.e., the available units in a non-probability sample. Therefore, the *missingness mechanism* (also referred to as *response mechanism*), which assumes a certain likelihood for every data point to be missing, becomes a *selection mechanism*. Furthermore, the three categories *missing completely at random* (MCAR, the likelihood of missing is the same for every data point), *missing at random* (MAR, the likelihood of missing is only the same within groups of measured auxiliary variables), *missing not at random* (MNAR, the likelihood for missing varies between data points for unknown/unmeasured reasons) become *selected completely at random* (SCAR), *selected at random* (SAR) and *selected not at random* (SNAR), respectively. While the original missingness mechanism only needs to focus on the missing observations in a design-based probability sample, the selection mechanism in a non-probability sample considers all observational units in the population.

For every inference problem an analyst needs to define which mechanism is assumed and is formalized through the model $P(S|Z, Y, \beta)$, where β corresponds to unknown parameters (Little and Rubin 2020). As indicated above for the three missingness mechanism categories, the strongest assumption, SCAR, defines $P(S|Z, Y, \beta) = P(S|\beta)$, meaning that S and $\{Y, Z\}$ are jointly independent. When assuming SCAR, β stands for the average selection rate and no selectivity correction for the sample needs to take place for an unbiased estimator. SAR, on the other hand, defines $P(S|Z, Y, \beta) = P(S|Z, \beta)$ and assumes that S and Y are conditionally independent given Z, i.e., the selectivity relies on the measured auxiliary variables and can therefore be completely identified. Finally, SNAR defines

$P(S|Z,Y,\beta) = P(S|Z,Y,\beta)$ and is the weakest assumption because it relies on all parameters defined for the selection mechanism. With SNAR the selection still depends on the unavailable data in Y, after conditioning on Z, making the complete degree of selectivity formally not identifiable from the data.

Essentially, most inference problems relying on non-probability samples need to assume either SAR or SNAR, as there is reason to believe that selected units will differ from unselected units. One needs to evaluate to what extent these differences can be detected through observed information (i.e., SAR) and to what extent these differences are undetectable with the data at hand (i.e., SNAR). Provided that one is able to compute the actual degree of selectivity (i.e., ground truth calibration), the resulting value can be decomposed into two parts, which ultimately relate to the difference between SAR and SNAR: (1) the ignorable part identified through SAR, and (2) the non-ignorable part characterized by SNAR. This is possible because the ignorable part can be fully identified. In practical cases, however, ground truth calibration is rarely possible, therefore, the only way to measure non-ignorable selectivity is to model the joint distribution $P(Y,S)$ as y_i and s_i are not assumed to be independent. Different strategies with additional (however unverifiable) assumptions can be used to model $P(Y,S)$ (Sikov 2018); one of them, namely pattern mixture modelling, will be introduced later.

While it has been established so far that by means of ground truth calibration distinction between the ignorable and non-ignorable part of selection bias is possible, another approach is to collect additional information on the non-ignorable part through historical information, if available. We hypothesize that proxy data, such as historical selectivity levels, offers additional insight in estimating $P(Y,S)$, especially when the time specific selectivity can be considered stable or follows identifiable seasonalities. Therefore, with regard to current indices for selectivity that rely on modelling the selection mechanism, the main contribution of this thesis project is the incorporation of additional information on historical selectivity levels, as well as to evaluate their performance.

Measuring Selectivity

In this Section, we will introduce two recently proposed measures that assess the impact of selection bias in inference problems, which are based on non-probability samples. After their introduction, the two measures are compared and adjusted to take historical selectivity information into account.

Standardized Measure of Unadjusted Bias (SMUB) by Little et al. (2020)

The first measure we discuss is given by Little et al. (2020) and is referred to as the standardized measure of unadjusted bias (SMUB). This measure assumes a normally distributed target variable Y and estimates the joint distribution $P(Y,S)$ by pattern mixture modelling (PMM) (Little 1994; Andridge and Little 2011). PMMs essentially split $P(Y,S)$ into two independent submodels according to S, leading to the isolation of the non-identifiable parameters from the identifiable ones. This is further explained in Sikov (2018, 421).

The SMUB is based on Maximum Likelihood (ML) estimates for a normal PMM relating X and Y, leading to the specification $P(Y,S) = P(Z,Y|S)$ (Little et al. 2020, 940). Complete identification of the submodels, $P(Z,Y|S=1)$ and $P(Z,Y|S=0)$, is therefore possible by assuming relations among the outcome distributions conditional on the selection mechanism. Each submodel is considered to follow a bivariate normal distribution with five parameters to be estimated (two means, two variances and a covariance) and the selection mechanism is constructed as follows:

$$P(S=1|Y,Z,\phi) = g(\phi Y + (1-\phi)Z), \tag{1}$$

where g is an arbitrary function ranging between (0,1) and ϕ is an unknown scalar parameter also ranging between (0,1). The parameter ϕ is the essential instrument in this selection mechanism as it regulates the degree of non-random selection after conditioning on Z (Little et al. 2020). According to Andridge and Little (2011), with this specification an ML estimate for $\bar{Y}$ can be expressed as a function of ϕ:

$$\bar{Y}(\phi) = \bar{y}^{(1)} + \frac{\phi + (1-\phi) r_{ZY}^{(1)}}{\phi r_{ZY}^{(1)} + (1-\phi)} \frac{\sqrt{\sigma^{2(1)}_{YY}}}{\sqrt{\sigma^{2(1)}_{ZZ}}} \left(\bar{Z} - \bar{z}^{(1)}\right), \tag{2}$$

where $\bar{Z}$ represents the population mean of the auxiliary variable. Conditioned on the selected units ($S = 1$), $\bar{y}^{(1)}$and $\bar{z}^{(1)}$ are the respective sample means, $\sigma^{2(1)}_{YY}$ and $\sigma^{2(1)}_{ZZ}$ stand for the respective sample variances and $r_{ZY}^{(1)}$ represents the sample correlation between Z and Y, which is transformed in a way that it ranges between (0,1). When rewriting this equality in terms of the deviation between the sample mean and the population mean $\left(\bar{Y} - \bar{y}^{(1)}\right)$, i.e., the bias, we receive the SMUB proposed by Little et al. (2020):

$$SMUB(\phi) = \frac{\phi + (1-\phi) r_{ZY}^{(1)}}{\phi r_{ZY}^{(1)} + (1-\phi)} * \frac{(\bar{z}^{(1)} - \bar{Z})}{\sqrt{s_{ZZ}^{(1)}}}. \tag{3}$$

When dissecting the right hand side of (3) into the two parts of the product, the first part represents a weighting adjustment (in the following “weight adjustment factor”), describing the interplay of ϕ and $r_{XY}^{(1)}$. The second part represents a standardized difference in means of the auxiliary variable between the selected and unselected population (in the following “standardized difference in means”). The magnitude of the latter is dependent on the individual research case, however, the weight adjustment factor can be identified within $(0, \infty)$ and is effectively minimized through high values for $r_{ZY}^{(1)}$.

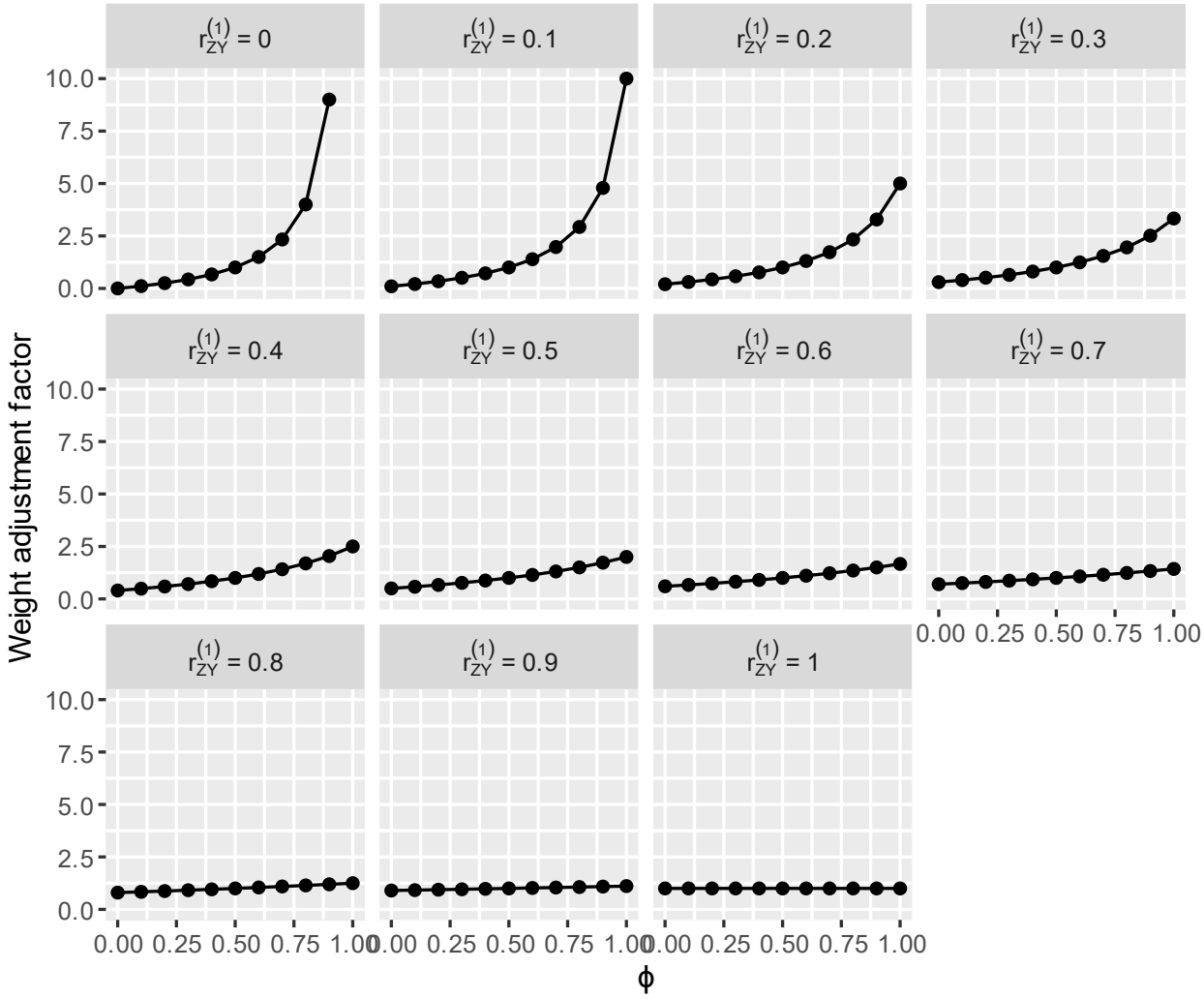

Figure 1 Theoretical *range of the weight adjustment factor given ϕ and $r_{ZY}^{(1)}$. The higher the sample correlation between the auxiliary variable Z and the target variable Y, i.e., $r_{ZY}^{(1)}$, the more stable the resulting weight adjustment factor – and ultimately – the bias estimate will be.*

Note: In the case of ϕ=1 and $r_{ZY}^{(1)} = 0$, zero-division takes place. This point has been excluded from the first panel.

Figure 1 presents the theoretical range of the weight adjustment factor for ordered values of $r_{ZY}^{(1)}$ and ϕ. Each panel shows the range given a particular value of $r_{ZY}^{(1)}$, the x-axis describes the given value of ϕ and the y-axis the resulting weight adjustment factor. The higher the correlation of Z and Y the more stable the weight adjustment factor across the whole range of ϕ – ultimately resulting in a more stable bias estimate. Furthermore, Little et al. (2020, 957) mention in their findings that the SMUB delivers acceptable results for the bias when $r_{ZY}^{(1)} > 0.4$.

However, by definition, the value for ϕ is inestimable from the sample. Little et al. (2020) suggest to calculate the range of possible SMUB values, using the values 0, 0.5 and 1 for a sensitivity analysis. Different values for ϕ suggest different selectivity mechanisms in the non-probability setting at hand. When $\phi = 0$, the selection mechanism is assumed to be SAR. When $\phi > 0$, part of the selectivity in the sample is non-ignorable (Rubin 1976), i.e., unbiased results can only be achieved when knowing the true values for ϕ and r_{ZY} (SNAR). Varying values for ϕ are especially necessary in settings where $r_{ZY}^{(1)} < 0.4$ because the stability of the weight adjustment factor decreases nearly exponentially. For example, when $r_{ZY}^{(1)} = 0.1$, the range of the weight adjustment factor is between (0,10). In this case, one has only access to an auxiliary variable that poorly correlates with the target variable. Here, it is especially relevant to find a narrower range for ϕ.

Given the mentioned assumptions, the prime advantage of the SMUB is that every parameter besides ϕ can be estimated from the available data.

Data Defect Index (d.d.i.) by Meng (2018)

The second approach to measure selection bias results from a different perspective on the topic delivered by Meng (2018).[2] According to Meng (2018), the bias of a finite population inference problem depends on three measures,

$$\bar{Y} - \bar{y}^{(1)} = \rho_{SY} \times \sigma_{YY} \times \sqrt{\frac{N-n}{n}}. \quad (4)$$

The first term describes the *data quality* ρ_{SY}, which represents the correlation between the selection mechanism and the target variable. The second term describes the *problem difficulty* σ_{YY}, which represents the standard deviation of the target variable, and the third term describes the *data quantity* $\sqrt{\frac{N-n}{n}}$, which represents the size of the sample in terms of the population size. These three terms

[2]The notation of Meng (2018) has been adjusted to follow suit in the paper at hand.

represent three error sources within any inference problem (i.e., probability or non-probability sample), which respectively can be minimized by either increasing the data quality, reducing the problem difficulty and/or increasing the data quantity. When using this equality with a high quality simple random sample (SRS), one can prove that $\mathrm{E}[\rho_{S,Y}] = 0$. This essentially describes the concept of selection bias because in an SRS without non-response we do not expect any correlation between the target variable and the selection-mechanism, i.e., SCAR. Any deviation from 0 quantifies a *data defect*, therefore, this can also be coined as a *data defect index* (d.d.i.), measuring the degree of selectivity in the particular inference problem (Meng 2018, 692).

Meng (2018, 714) computes the bias of the USA 2016 election polls after the election took place by viewing the actual results as $\bar{Y}$. He solves for $\rho_{S,Y}$ and shows that "[…] there is a consistent pattern of underreporting by Trump supporters, inducing on average about -0.005 data defect correlation" (Meng 2018, 714). He suggests using this value as an estimate for $\hat{\rho}_{SY}$ in subsequent election polls (e.g., in 2020) to correct for selectivity. Therefore, Meng's example (2018) applies to recurring data collection settings where the actual value of $\rho_{S,Y}$ can be assessed and builds on the assumption that the selection mechanism does not change over time.

The exact value for d.d.i. in (4) does not rely on any assumptions, except that the population is known. However, this limits its practical use severely, as only one of the three terms in (4), namely $\sqrt{\frac{N-n}{n}}$, can be derived from the non-probability sample at hand. ρ_{SY} and σ_{YY} are expressed on the population level and cannot be derived from the sample.

Comparing the SMUB and the d.d.i.

We focus on these two estimation strategies because, according to their respective authors, they provide promising performance for estimating bias in non-probability samples. This needs to be underlined for the SMUB as Boonstra et al. (2019) show, in an updated simulation study originally from Nishimura et al. (2016), comparing different indices for measuring selection bias, that the SMUB is most

predictive of the true extent of selection bias. Furthermore, to the best of the authors knowledge, the SMUB and the d.d.i. were not compared to each other formally. This Section is dedicated to fill this gap.

Both approaches deliver a framework to quantify the selection bias of an inference problem ($\bar{Y} - \bar{y}^{(1)}$). However, the SMUB offers a much more practical workflow as all parameters, except for ϕ, can be estimated from the non-probability sample at hand, whereas for the d.d.i. it is largely unclear how to estimate the necessary parameters.

But how do they actually relate to each other? Both measures rely on the concept that the selection mechanism is responsible for the potential lack of representativity in the sample. In case of SAR, this means that the selection mechanism can be modeled with the use of auxiliary variables. The d.d.i. formally does not rely on auxiliary variables, it relies on S. This means when we use S instead of Z, ϕ in (1) becomes equal to zero and g becomes equal to I, the indicator function (i.e., the function such that $I(Z) = 1$ if $Z = 1$ and $I(Z) = 0$ if $Z = 0$). Thus, (1) with this specification states $\Pr(S) = I(S)$.

This implies that the weight adjustment factor in the SMUB can be reduced to

$$\frac{\phi + (1-\phi)r_{ZY}^{(1)}}{\phi r_{ZY}^{(1)} + (1-\phi)} = \frac{\phi + (1-\phi)r_{SY}^{(1)}}{\phi r_{SY}^{(1)} + (1-\phi)} = r_{SY}^{(1)}. \tag{5}$$

In terms of Meng (2018) and (4), $r_{SY}^{(1)}$ is denoted as ρ_{SY}. Now, according to (2), we can write the bias of $\bar{y}$ as

$$\bar{Y} - \bar{y}^{(1)} = -r_{SY}^{(1)} \sqrt{\frac{{\sigma^2}_{YY}^{(1)}}{{\sigma^2}_{SS}^{(1)}}} \left(\bar{S} - \bar{s}^{(1)}\right). \tag{6}$$

As mentioned above, the SMUB tries to derive an expression for the bias of the mean of a target variable from the perspective where all relevant terms must be estimated from the available (selective) dataset. In an ideal world, one would not have to estimate $\frac{{\sigma^2}_{YY}^{(1)}}{{\sigma^2}_{SS}^{(1)}}$, but rather use $\frac{{\sigma^2}_{YY}}{{\sigma^2}_{SS}}$ directly, i.e., the population variance of Y divided by the population variance of S. This is the perspective of Meng (2018).

For a general auxiliary variable Z, in many cases $\frac{\sigma^2{}^{(1)}_{YY}}{\sigma^2{}^{(1)}_{SS}}$ may be a good estimate for $\frac{\sigma^2{}_{YY}}{\sigma^2{}_{SS}}$. This is, however, not the case when $Z = S$. When $Z = S$, $\sigma^2{}^{(1)}_{ZZ} = \sigma^2{}^{(1)}_{SS} = 0$, whereas $\sigma^2{}_{SS} \neq 0$. Fortunately, calculating the population variance $\sigma^2{}_{SS}$ of S is quite simple. The selection process is a binomial process, where a unit is selected from N units in total and n units are selected. Each of the n units selected in the sample have a value of 1 on S. The $N - n$ units that are not selected in the sample each have a value of 0 on S. Thus, $\sigma^2{}_{SS} = \frac{n}{N}\left(1 - \frac{n}{N}\right)$, i.e., the variance is simply $p(1-p)$, where $p = \frac{n}{N}$. Consequently, $\sqrt{\sigma^2{}_{SS}} = \sqrt{\frac{n}{N}\left(1 - \frac{n}{N}\right)}$.

Furthermore, $\bar{S} - \bar{s}^{(1)} = \frac{n}{N} - 1 = -\left(1 - \frac{n}{N}\right)$, since $s^{(1)} = 1$ for all units i that are selected in the sample ($\bar{s}^{(1)} = 1$), and $S = 1$ for n out of N units in the population (namely the units that are selected in the sample) and $S = 0$ otherwise ($\bar{S} = \frac{n}{N}$).

Combining all terms and comparing (6) with (4) we get:

$$\bar{Y} - \bar{y}^{(1)} = -r^{(1)}_{SY}\sqrt{\frac{\sigma^2{}^{(1)}_{YY}}{\sigma^2{}^{(1)}_{SS}}}\left(\bar{S} - \bar{s}^{(1)}\right) = -r^{(1)}_{SY}\sqrt{\sigma^2{}_{YY}}\frac{\bar{S} - \bar{s}^{(1)}}{\sqrt{\sigma^2{}_{SS}}} = \rho_{SY}\sigma_{YY}\frac{1 - \frac{n}{N}}{\sqrt{\frac{n}{N}\left(1 - \frac{n}{N}\right)}}$$

$$= \rho_{SY} \times \sigma_{YY} \times \sqrt{\frac{N-n}{n}}. \qquad (7)$$

In conclusion, it can be said that formula (4) is a special case of formula (2), namely when one uses the selection indicator as the predictor in the selection mechanism. This underlines the more general usage of the SMUB in comparison to the d.d.i. As mentioned above, if $\phi = 0$ the selection mechanism is SAR and unbiased estimations of $\bar{Y}$ are possible. The fact that ϕ needs to be zero in the d.d.i. implies that the d.d.i. cannot handle a SNAR assumption; ϕ, which, as mentioned above, regulates the degree of non-random selection after conditioning on Z, is basically dropped because we have access to S. No part of the selection bias is left after conditioning the selection mechanism on S. This is illustrated through the

standardized measure of *adjusted* bias (SMAB) (Little et al. 2020, 942), $SMAB(\phi) = SMUB(\phi) - SMUB(0)$. This expression basically computes the difference between a SAR assumption ($SMUB(0)$) and a SNAR assumption ($SMUB(\phi)$). For the true value of ϕ, this expression results in the non-ignorable part of the selection bias in the specified inference problem at hand. Therefore, the SMUB has the ability to decompose the total selection bias into the ignorable and non-ignorable part, always depending on which auxiliary variables are specified in the selection mechanism.

Utilizing historical information for the estimation of parameters

We have established that the SMUB is a valid measure for selection bias estimation and decomposition, if we have access to the true value of ϕ. When this is not possible, as in most applied cases, Little et al. (2020) suggest to perform a sensitivity analysis for the whole range of ϕ. As mentioned above, time-series inference problems contain additional information through their repeated nature. Meng (2018) proposes in his measure to use such proxy data, e.g., historical information, as additional input for the estimation of the necessary parameters.

Therefore, we formulate the following research question: Concerning time-series inference problems, how accurately/reliably can the selection bias be estimated when one derives the necessary parameters for the SMUB and the d.d.i. from proxy data such as historical selectivity levels?

For this we add to the assumptions made in Section 2. Suppose an inference problem that is based on $\boldsymbol{D} = \{y_i^k, z_i^k, i = 1, \dots, n\}$, where the additional superscript k denotes the time period of a specified time-series (e.g., years, quarters) with $k = 1, \dots , T$ time periods. $\bar{Y}^T$ indicates the average outcome of the target variable for the most recent (focal) time period, which is therefore the primary estimand of interest. The units in the population U stay the same across time periods, but the respective values for Y and Z can vary with k. Furthermore, we assume that we have access to the true values of $\bar{Y}^{T-k}$ (previous time periods) as these can be computed retrospectively.

For the SMUB only one component, namely ϕ^t, needs to be estimated and for the d.d.i. two components, σ_Y^t and $\rho_{S,Y}^t$, need to be estimated. The respective remaining components of the measures can be derived from the data. For each component that needs to be estimated different estimators will be formalized, which utilize historical information as mentioned above.

We suggest four estimation strategies for incorporating historical information. First, the *Stability-hypothesis*: using the estimate from the previous period $(T-1)$. Second, the *Seasonality hypothesis*: assuming periods are repeated in defined cycles, such as quarters in years, we can use the estimate from the same period a full year ago (here, $T-4$). Third, the *Extended stability hypothesis:* using the (weighted) average of the previous periods' estimates. Fourth, the *Combined hypothesis*: using a (weighted) average of the previous three estimation strategies. These strategies rely on the assumption that estimates are not independent over time.

It should be noted that the work on this Section is still under development and the following topics will be integrated for the final thesis. First, further ideas on incorporating historical information will be added, focusing especially on research concerning multivariate PMMs, which can be found in Little (2009) and Little (1996). Essentially, these models can handle multiple target variables, which might provide a developed framework for handling time-series inference problems – this is still under scrutiny. Subsequently, the concrete estimators for the adjusted measures can be derived from the proposed strategies and formalized.

Case Study

The final section of this research report devotes its attention to the evaluation-plan of the adjusted indices for measuring selection bias in time-series inference problems. We have decided against a simulation study and for a real-life case study for the following reason. Time-series inference problems that are based on non-probability samples are a special type of inference problems. They contain additional information that is not available in general inference problems, such as historical selectivity levels. Therefore, we can surely expect performance increases through the incorporation of historical selectivity levels compared to, for example, the currently suggested workflow for the SMUB, namely conducting a sensitivity analysis for the whole range of ϕ (Little et al. 2020). However, there are no extensions yet that focus on optimally estimating ϕ. These extensions are very important because they can highlight the flexibility of the indices in incorporating problem-specific hypotheses concerning parameter estimation. After all, potential advantages of historical information are always dependent on the inference problem-specific relations in the target variable between periods, i.e., the hypotheses mentioned in the previous Section. Therefore, we expect the scientific community to benefit more from a practical execution of our adjusted measures of the SMUB and the d.d.i.

Particularly in the realm of Official Statistics, inference problems are often repeated for certain time periods such as years or quarters. Here, reliable and accurate estimates are of great importance as they form the basis for policy making. Current developments in this research field have shown that increasingly more national statistical institutes are shifting from surveys to using administrative data sources, such as registry data, for the production of Official Statistics. One reason for this are increasing non-response rates in surveys, decreasing their quality (De Broe et al. 2020). Registry data allows for (almost) complete data access on all units in the population, offering the potential for accurate estimations. One example are short term business statistics (STS), which are computed on a quarter yearly basis and can rely on value added tax (VAT) registry data (Ouwehand and Schouten 2014; Eurostat 2002).

The main goal of STS is the provision of early indicators describing the economic activity, i.e., they provide predictions. In this case, selectivity might be introduced because period specific predictions are necessary before the period ends and all units in the registry have provided their data. This selectivity needs to be corrected for as good as possible. Once the complete data is available – usually at the end of the focal period – the actual level of bias of the estimator, based on the available data for Y, can be retrospectively computed for any time point in the period.

For the case study we focus on the estimation of the average turnover (target variable) in a given period. For this we use Dutch VAT data for the years 2014-2016, resulting in twelve quarter yearly time periods. Access has been provided by Statistics Netherlands and ethical consent by Utrecht University for the following relevant variables:

- Business ID (anonymized)
- NACE category
- Turnover
- Number of employees
- Quarter
- Receiving date of the data

VAT data can be considered as a gradually filling data base, which is indicated through the "Receiving date of the data" variable. Each day in any period presents an updated cumulative sample of the register, i.e., each day represents a new non-probability sample of the population with varying degrees of selectivity. As an auxiliary variable for the selection mechanism we will use the "Number of employees" variable, which shows acceptable levels of correlation with the target variable (Ouwehand and Schouten 2014).

As in the previous Section, this Section is still under development, as the complete design of the case study is not final yet. This relates mostly to the formalized performance evaluation of the adjusted indices. We have discussed so far to evaluate the performance by splitting the samples according to NACE

categories. This variable describes the type of economic activity of a business. The advantage of this is to create multiple samples of varying degrees of selectivity, which can help in measuring the performance. Simultaneously, such a splitting strategy would also constitutes a viable approach in tackling the inference problem – highlighting the practical showcase of the case study.

References

Andridge, Rebecca, and Roderick Little. 2011. “Proxy Pattern-Mixture Analysis for Survey Nonresponse.” *Journal of Official Statistics* 27 (2): 153–80.

Boonstra, Philip, Roderick Little, Brady West, Rebecca Andridge, and Fernanda Alvarado-Leiton. 2019. “A Simulation Study of Diagnostics for Bias in Non-Probability Samples.” 125.

Brick, J. M., and D. Williams. 2013. “Explaining Rising Nonresponse Rates in Cross-Sectional Surveys.” *The Annals of the American Academy of Political and Social Science* 645 (1): 36–59. https://doi.org/10.1177/0002716212456834.

Broe, Sofie De, Peter Struijs, Piet Daas, Arnout van Delden, Joep Burger, Jan van den Brakel, Olav ten Bosch, Kees Zeelenberg, and Winfried Ypma. 2020. “Updating the Paradigm of Official Statistics: New Quality Criteria for Integrating New Data and Methods in Official Statistics.” 02–20.

Callegaro, M, A Villar, D Yeager, and J Krosnick. 2014. “A Critical Review of Studies Investigating the Quality of Data Obtained with Online Panels Based on Probability and Nonprobability Samples.” In *Online Panel Research: A Data Quality Perspective*, edited by M Callegaro, R Baker, J Bethlehem, A Göritz, and J Krosnick, 23–53. UK: John Wiley and Sons.

Cornesse, Carina, Annelies G. Blom, David Dutwin, Jon A. Krosnick, Edith D. De Leeuw, Stéphane Legleye, Josh Pasek, et al. 2020. “A Review of Conceptual Approaches and Empirical Evidence on Probability and Nonprobability Sample Survey Research.” *Journal of Survey Statistics and Methodology* 8 (1): 4–36. https://doi.org/10.1093/jssam/smz041.

Eurostat. 2002. *Methodology of Short-Term Business Statistics: Interpretation and Guidelines*.

Kalton, Graham. 2019. “Developments in Survey Research over the Past 60 Years: A Personal Perspective.” *International Statistical Review* 87 (S1): S10–30. https://doi.org/10.1111/insr.12287.

Little, Roderick. 1994. “A Class of Pattern-Mixture Models for Normal Incomplete Data.” *Biometrika* 81 (3): 471–83. https://doi.org/10.1093/biomet/81.3.471.

———. 1996. “Pattern-Mixture Models for Multivariate Incomplete Data with Covariates.” *Biometrics* 52 (1): 98. https://doi.org/10.2307/2533148.

———. 2009. “Selection and Pattern-Mixture Models.” In *Longitudinal Data Analysis*, edited by G.M Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs. Boca Raton, FL: CRC Press.

Little, Roderick, and Donald Rubin. 2020. *Statistical Analysis with Missing Data*. 3rd ed. Hoboken, NJ: JohnWiley & Sons, Inc. https://doi.org/10.2307/2982783.

Little, Roderick, Brady West, Philip Boonstra, and Jingwei Hu. 2020. “Measures of the Degree of Departure from Ignorable Sample Selection.” *Journal of Survey Statistics and Methodology* 5 (8): 932–64. https://doi.org/10.1093/jssam/smz023.

Meng, Xiao Li. 2018. “Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations, Big Data Paradox, and the 2016 Us Presidential Election.” *Annals of Applied Statistics* 12 (2): 685–726. https://doi.org/10.1214/18-AOAS1161SF.

Nishimura, Raphael, James Wagner, and Michael Elliott. 2016. “Alternative Indicators for the Risk of Non-Response Bias: A Simulation Study.” *International Statistical Review* 84 (1): 43–62. https://doi.org/10.1111/insr.12100.

Ouwehand, Pim, and Barry Schouten. 2014. “Measuring Representativeness of Short-Term Business Statistics.” *Journal of Official Statistics* 30 (4): 623–49. https://doi.org/10.2478/JOS-2014-0041.

Rubin, Donald. 1976. “Inference and Missing Data.” *Biometrika* 63 (3): 581–92.

Sikov, Anna. 2018. “A Brief Review of Approaches to Non-Ignorable Non-Response.” *International Statistical Review* 86 (3): 415–41. https://doi.org/10.1111/insr.12264.

Valliant, Richard. 2020. “Comparing Alternatives for Estimation from Nonprobability Samples.” *Journal of Survey Statistics and Methodology* 8 (2): 231–63. https://doi.org/10.1093/jssam/smz003.