Deliverable Number: 2.8

Deliverable Title: Report on existing approaches to computing calibrated weights and possible improvements

Work Package: 2 - Representing the population

Deliverable type: Report

Dissemination status: Public

Submitted by: SHARE ERIC

Authors: Giuseppe De Luca (MEA)

Date Submitted: December, 2016

www.seriss.eu @SERISS_EU

SERISS (Synergies for Europe's Research Infrastructures in the Social Sciences) aims to exploit synergies, foster collaboration and develop shared standards between Europe's social science infrastructures in order to better equip these infrastructures to play a major role in addressing Europe's grand societal challenges and ensure that European policymaking is built on a solid base of the highest-quality socio-economic evidence.

The four year project (2015-19) is a collaboration between the three leading European Research Infrastructures in the social sciences – the European Social Survey (ESS ERIC), the Survey for Health Aging and Retirement in Europe (SHARE ERIC) and the Consortium of European Social Science Data Archives (CESSDA AS) – and organisations representing the Generations and Gender Programme (GGP), European Values Study (EVS) and the WageIndicator Survey.

Work focuses on three key areas: Addressing key challenges for cross-national data collection, breaking down barriers between social science infrastructures and embracing the future of the social sciences.

Please cite this deliverable as: De Luca, G. (2016) *Report on existing approaches to computing calibrated weights and possible improvements*. Deliverable 2.8 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

Report on existing approaches to computing calibrated weights and possible improvements

Giuseppe De Luca
University of Palermo, Italy

December 20, 2016

Abstract

In this report we review two broad classes of weighting methods to compensate for nonresponse errors in sample surveys: the nonresponse calibration approach and the propensity score approach. We first show that arbitrary choices of the distance function characterizing the calibration methodology correspond to assuming, at least implicitly, alternative parametric models for the nonresponse process. As a natural extension of the nonresponse calibration approach, we then introduce the propensity score approach that allows us to improve the robustness of the survey weights by estimating an explicit model for the nonresponse process. Since the choice between these two approaches is not always clear-cut, we also consider a two-step procedure which involves a calibration adjustment in the first step and a propensity score adjustment in the second stage. Finally, we conclude our report with a discussion on the important role played by the auxiliary information which is available to compensate for nonresponse errors.

Contents

1 Introduction **2**

2 Calibration approaches **3**
2.1 Calibration under complete response 3
2.2 Nonresponse calibration 5

3 Propensity score approaches **7**

4 Auxiliary variables **9**

5 Conclusions **11**

References **14**

1 Introduction

Weighting adjustment methods are commonly employed in sample surveys to compensate for both discrepancies of the realized weighted-sample estimates from known population values and other systematic sources of nonsampling errors such as undercoverage, unit nonresponse, and panel attrition. In this report, we review the current state of the survey literature on this important topic to emphasize advantages and weaknesses of the methods in widespread current use.

We focus throughout on unit nonresponse, which is defined as the failure to interview a sample unit because of either noncontact or explicit refusal to participate in a cross-sectional survey (or, equivalently, the first wave of a longitudinal survey). As other types of nonresponse errors, unit nonresponse may increase the mean squared error (MSE) of sample-based estimators through two channels. By reducing the available sample size, nonresponse leads necessarily to a loss of precision (larger sampling variance) in estimating the population parameters of interest. This loss of precision does not need to be the main concern, however, as systematic differences between respondents and nonrespondents may also lead to biased estimators. Although the finite-sample precision of the estimators is important, the greater concern on the bias is typically justified by the fact this component of the MSE does not vanish as the sample size increases.

We shall discuss two broad classes of weighting methods. The first one is the calibration approach introduced by Deville and Särndal (1992) which encompasses as special cases many traditional weighting procedures such as poststratification, raking, and generalized linear regression. Following its original formulation, we first describe the key features of the calibration methodology by focusing on an ideal setting with complete response. This choice is motivated by the fact the practical implementation of the calibration approach can be easily generalized to account for unit nonresponse errors. Not surprisingly, however, the most relevant implication of this generalization is on the statistical properties of the nonresponse calibration estimators. The most striking issue is that arbitrary choices of the distance function for the calibration methodology correspond to assuming, at least implicitly, alternative parametric models for the nonresponse process.

This limitation of the calibration approach leads us to introduce the class of propensity score approaches which aims to improve the robustness of the survey weights by estimating an explicit model for the nonresponse process. Here, we discuss several issues including parametric and nonparametric estimation procedures, doubly robust weighting methods for handling model misspecification problems, as well as other recent developments that allow tailoring the standard propensity score approach to specific missing data patterns and more general nonresponse mechanisms. We shall

also discuss briefly the so-called two-step procedures that combine propensity score and calibration adjustments to the survey weights in two sequential steps.

After reviewing the available weighting methods, we put some emphasis on the critical - but also fascinating - issue regarding the role of auxiliary information when adjusting survey weights for nonresponse errors. Specifically, we discuss briefly the different types of auxiliary variables that can be used in the calibration and the propensity score approaches and the different types of information that is typical available to study the various sources of nonresponse errors. For cases when there exists a large set of auxiliary variables, we also discuss the importance of handling the bias-precision trade-off surrounding the choice of the auxiliary variables to be used in the construction of the survey weights.

2 Calibration approaches

The calibration approach by Deville and Särndal (1992) is a procedure to adjust the survey weights so that the weighted sum of a vector of benchmark variables over the sample units equals the corresponding vector of population totals. The rationale behind this procedure is that by ensuring consistency between the sample and the population distributions of some benchmark variables, the calibrated weights will also perform well when applied to the study variables of interest. In Section 2.1 we first discuss the key features of the calibration methodology by focusing on an ideal setting with complete response. This unrealistic framework is extended in Section 2.2, where we discuss the so-called nonresponse calibration approach for handing unit nonresponse errors.

2.1 Calibration under complete response

Let us consider a finite population $U = \{1, \ldots, i, \ldots, N\}$ of N elements, from which a probability sample $s = \{1, \ldots, i, \ldots, n\} \subseteq U$ of size $n \leq N$ is drawn according to the sampling design $p(\cdot)$. Unless otherwise specified, we shall assume that the inclusion probability $\pi_i = \Pr(i \in s)$ and the resulting *design weights* $w_i = \pi_i^{-1}$ are known and strictly positive for all population units. The availability of the design weights w_i allows us to account for the randomness due to probability sampling. For example, if we wish to estimate the population total $t_y = \sum_{i \in U} y_i$ of a study variable y, then the Horvitz-Thompson estimator

$$\widehat{t}_y = \sum_{i \in s} w_i y_i \tag{1}$$

is known to be design unbiased, that is $\mathbb{E}_p(\widehat{t}_y) = t_y$, where $\mathbb{E}_p(\cdot)$ denotes the expectation with respect to the sampling design.

Let us assume next that additional information is available to construct a class of more efficient estimators. More precisely, let $x_i = (x_{i1}, \ldots, x_{iq})^\top$ be a q-vector of auxiliary variables for which we known the corresponding population totals $t_x = \sum_{i \in U} x_i$ from either the sampling frame or other external sources such as census data and administrative archives. We shall refer to the auxiliary variables x_i as calibration variables and to the population totals t_x as U-level calibration margins. The basic idea of the calibration approach is to determine a new set of *calibrated weights* w_i^* that are close as possible (in an average sense with respect to a given distance function) to the design weights w_i, while also satisfying the constraints

$$\sum_{i \in s} w_i^* x_i = t_x. \tag{2}$$

Thus, given a distance function $G(w_i^*, w_i)$, calibration consists of minimizing the overall distance $\sum_{i \in s} G(w_i^*, w_i)$ with respect to w_i^* subject to a set of equality constraints. Deville and Särndal (1992) show that, under mild regularity conditions, the solution of this constrained minimization problem gives

$$w_i^* = w_i F(\eta_i), \qquad i = 1, \ldots, n, \tag{3}$$

where $\eta_i = x_i^\top \lambda$ is a linear combination of the calibration variables x_i, λ is a q-vector of Lagrangian multipliers associated with the constraints (2), and $F(\cdot)$ is a monotonic and twice-differentiable calibration function that is uniquely related to the distance function $G(\cdot, \cdot)$ and satisfies the restrictions $F(0) = 1$ and $F'(0) = 1$. Equation (3) shows that the ratio between the calibrated weights w_i^* and the design weights w_i depends on the calibration function $F(\cdot)$ (or, equivalently, the distance function $G(\cdot, \cdot)$) and the vector x_i of calibration variables. A distinguishing feature of this approach is that many traditional re-weighting procedure such as post-stratification, raking, and generalized linear regression (GREG) can be viewed as special cases of calibration estimator

$$\widehat{t}_y^* = \sum_{i \in s} w_i^* y_i, \tag{4}$$

for particular choices of the calibration function $F(\cdot)$ and the vector of calibration variables x_i.

Popular specifications of the calibration function are the linear form $F(u) = 1 + u$, the exponential form $F(u) = \exp(u)$, the truncated-linear form $F(u) = \min\{M, \max\{L, 1 + u\}\}$, and the logit form

$$F(u) = \frac{L(M-1) + M(1-L)\exp(Au)}{M - 1 + (1-L)\exp(Au)},$$

where L and M denote predefined lower and upper bounds, respectively, and

$$A = \frac{M - L}{(1 - L)(M - 1)}.$$

The linear specification, which derives from a chi-quare distance function and leads to the widely used GREG estimator, has the advantage of ensuring a closed form solution for the calibrated weights w_i^*. Depending on the chosen set of calibration variables, the resulting weights can be however negative or extremely large because the underlying distance function is unbounded. The other specifications of the calibration function allow to avoid these issues, but a solution for the calibration problem may not exist and the computation of the associated Lagrangian multipliers usually requires iterative techniques. In particular, the exponential specification avoids the problem of negative weights, but it is still subject to the issue of large variability of the weights that in turn affects the precision of the calibration estimator $\widehat{t}_y^*$. The truncated-linear and logit specifications are usually preferred to the other specifications because they allow to restrict in advance the range of feasible values for the calibrated weights by suitable choices of the lower bound L and the upper bound M.

As pointed out by Deville and Särndal (1992), effectiveness of the calibrated weights depends crucially on the correlation between the study variable y and the calibration variables x. In the extreme case when y can be expressed as a linear combination of x the calibrated estimator $\hat{t}_y^*$ gives an exact estimate of t_y for every realized sample s. Deville and Särndal (1992) also show that, under suitable regularity conditions, the calibration estimator $\hat{t}_y^*$ has desirable asymptotic properties. Moreover, all calibration estimators are asymptotically equivalent to the GREG estimator resulting from a linear specification of the calibration function. Thus, in large samples, the calibrated weights are robust to arbitrary choices of the functional form for $F(\cdot)$. Unfortunately, this robustness property does not extend to the more realistic setting where survey data are affected by nonresponse errors (see, for example, Haziza and Lesage 2016). We shall expand on this issue in the following section.

2.2 Nonresponse calibration

The calibration approach can be easily generalized to relax the strong assumption of complete response. Following Lundström and Särndal (1999) and Särndal and Lundström (2005), we shall refer to this generalization as nonresponse calibration, but it is also know as the one-step approach to emphasize the distinction with other two-step approaches where explicit adjustments for nonresponse errors are performed in a preliminary stage before applying calibration adjustments.

To extend the complete response setting we now assume that only a subsample $s_r \subseteq s$ of $n_r \leq n$ units agree to participate into the survey. Here, the standard justification for calibration treats survey response as an additional phase of the sampling design (see, e.g., Politz and Simmons 1949; Hartley 1946; Oh and Scheuren 1983). Survey response is therefore a random outcome and we shall denote by ϕ_i, $i = 1, \ldots, n$, the response propensity of the ith sample unit. As before, calibration consists of finding a set of nonresponse calibrated weights $\widetilde{w}_i^*$, $i = 1, \ldots, n_r$, that are as close as possible to the design weights w_i, while also respecting the calibration equations

$$\sum_{i \in s_r} \widetilde{w}_i^* x_i = t_x. \tag{5}$$

The solution of this constrained minimization problem gives *nonresponse calibrated weights* of the form

$$\widetilde{w}_i^* = w_i F(x_i^\top \lambda), \qquad i = 1, \ldots, n_r, \tag{6}$$

for many alternative specifications of the calibration function $F(\cdot)$ and different choices of calibration variables x_i. The class of nonresponse calibrated estimators of the population totals t_y is then given by

$$\widetilde{t}_y^* = \sum_{i \in s_r} \widetilde{w}_i^* y_i. \tag{7}$$

At first glance, there are few differences with respect to the complete response setting. The calibration procedure is now restricted to the complete-case data $\{(y_i, x_i, w_i) : i \in s_r\}$ rather than to the complete data $\{(y_i, x_i, w_i) : i \in s\}$, but its key features are essentially the same. However, upon reflection, we realize that the statistical properties of the two types of calibration estimators can be substantially different because of the additional randomness due to the nonresponse mechanism. Lundström and Särndal (1999) give expressions for the bias, the variance and the MSE of the GREG estimator which is a special case of the nonresponse calibration estimator $\widetilde{t}_y^*$ when $F(\cdot)$ has a linear specification. A more general expression for the bias of the whole class of nonresponse calibrated estimators can be found in Haziza and Lesage (2016). These studies show that, unlike the complete response setting, there exists now two set of conditions under which $\widetilde{t}_y^*$ is an (approximately) unbiased estimator of t_y: (i) the nonresponse mechanism is missing at random (MAR - Rubin 1976) and $y_i = x_i^\top \beta + \epsilon_i$ with $\mathbb{E}(\epsilon_i | x_i) = 0$; and (ii) $F_i = \phi_i^{-1}$. Condition (i) can be viewed as a natural extension of the unbiasedness property when survey data suffer from nonresponse errors. Condition (ii) is more interesting and perhaps surprising as it emphasizes that, even though the calibration approach does not require an explicit model for the nonresponse mechanism, alternative specifications of the calibration function $F(\cdot)$ correspond in practice to different parametric

models for the relationship between the response propensity ϕ_i and the calibration variables x_i. In some sense, assumptions about the nonresponse mechanism are implicit in the specification of the calibration function $F(\cdot)$ and the misspecification of this functional form may now lead to biased nonresponse calibrated estimators.

3 Propensity score approaches

Our previous considerations of the nonresponse calibration estimator lead us to introduce another wide class of weighting methods, known as propensity score approaches, that involve some explicit model for the nonresponse process. The theoretical framework of this approach was originally introduced by Rosenbaum and Rubin (1983) in the context of observational studies for estimating the average causal effects of a treatment whose assignment is not randomized. A useful overview of the propensity score approach for nonresponse in sample surveys can be found in Brick (2013). The basic idea is to take advantage of the dimension-reducing property of the propensity score, which is defined as the conditional probability of participating in the survey given covariates, to restate the missing at random (MAR) assumption as

$$\Pr\{R_i = 1|y_i, z_i\} = \Pr\{R_i = 1|z_i\} = \phi_i, \tag{8}$$

where R_i is a binary response indicator that takes value one if the ith sample unit participates in the survey and value zero otherwise, $\phi_i = \phi(z_i)$ is the true response propensity, and z_i is a k-vector of auxiliary variables that are observed for all sample units $i \in s$ (both respondents and nonrespondents). Notice that a comparative advantage with respect to the nonresponse calibration approach is that estimation of the unknown $\phi_i = \phi(z_i)$ can now exploit auxiliary variables of various types (both discrete and continuous).

Given a consistent estimate $\widehat{\phi}_i$ of the response propensity ϕ_i, corrections for nonresponse errors can be based on a variety of propensity score approaches. The most simple strategy consists of using the *propensity score weights* $\widehat{w}_i = w_i/\widehat{\phi}_i$ to directly compute a propensity-score adjusted (PSA) estimator of t_y

$$\widehat{t}_y = \sum_{i \in s_r} \widehat{w}_i y_i. \tag{9}$$

Since the propensity score weights $\widehat{w}_i$ do not have to satisfy the calibration equations associated with a set of benchmark variables, it may be reasonable to apply an additional calibration adjustment to obtain the so-called *propensity score calibrated weights* $\widehat{w}_i^*$. This leads to a two-step procedure which involves a propensity score adjustment in the first step and a calibration adjustment in the

second step. Notice that, by using $\widehat{w}_i$ as basis to compute $\widehat{w}_i^*$, we can now calibrate against either a set of U-level calibration margins

$$\sum_{i \in s_r} \widehat{w}_i^* x_i = t_x \tag{10}$$

or a set of S-level calibration margins

$$\sum_{i \in s_r} \widehat{w}_i^* x_i = \sum_{i \in s} \widehat{w}_i x_i. \tag{11}$$

The latter set of constraints requires that the auxiliary variables x_i are observed for all sample units, but it allows us to replace the possibly lacking information on the population totals t_x with an unbiased sample estimate. A third approach is the so-called "weighting class approach" (see, e.g., Little 1986; Sarndal et al. 1992; Eltinge and Yanaseh 1997; Haziza and Beaumont 2007), another widely used two-step procedure which first partitions the subsample of respondents into weighting classes on the basis of $\widehat{\phi}_i$ and then adjusts the design weights on the basis of the within-class response rate.

A common requirement of these alternative variants of the propensity score approach is the availability of a consistent estimator of the response propensity. The standard approach to estimation relies on the parametric maximum likelihood estimator of a logit/probit model for the binary indicator R_i given the auxiliary variables z_i (Ekholm and Laaksonen 1991). If the assumed parametric model is correctly specified, then the resulting PSA estimator $\widehat{t}_y$ is unbiased and consistent for t_y (see, e.g., Kim and Kim 2007). However, incorrect specification of the underlying distributional assumptions leads to inconsistent estimates. This explains why nonparametric methods based on kernel, local polynomial, and regression tree estimators are typically preferred to standard maximum likelihood estimators (see, e.g., Da Silva 2003; Da Silva and Opsomer 2006, 2009; Phipps and Toth 2012). In some cases, inconsistency may also be due to a failure of the assumption that response propensities are independent across sample units (Skinner and D'Arrigo 2011; Brick 2013). This problem may arise, for example, in sample surveys based on a multistage design for the purpose of reducing travel costs in face-to-face interviews and those designed to interview more than one person per household. In these cases, consistent estimation of the response propensities requires to account for the possible correlation in the response behaviors of the sample units by using nonparametric methods for multivariate response models.

The strong assumption of a correctly specified model for the response propensities can also be relaxed with the aid of doubly robust weighting procedures that involve the specification of two models, one for response propensities and one for the conditional distribution of the study variables

given the auxiliary variables. The attractive feature of these doubly robust weighting procedures is that they provide consistent estimators of the population parameters whenever one, but not necessarily both, of the two models has been correctly specified. Recent overviews of doubly robust methods can be found in Hu et al. (2012), Han (2014), Rotnitzky and Vansteelandt (2015), and Vermeulen and Vansteelandt (2015).

Another fruitful line of research has been concerned with the issue of tailoring the standard propensity score approach to specific missing data patterns and more general response mechanisms. For example, Li et al. (2013) discuss propensity score approaches under three alternative missing data patterns (uniform missingness, monotone missingness and non-monotone missingness) and two alternative nonresponse mechanisms (MAR and not missing at random - NMAR). Interestingly, they show that weighting methods under MAR mechanisms can be naturally extended to MNAR mechanisms through the specification of a selection-bias function that quantifies the residual association of the missing probabilities and the unobserved data after adjusting for the observed data. In the same spirit, Robins and co-authors have proposed a randomized monotone missingness approach to analyze non-monotone ignorable missing data (Robins et al. 1997) and a selection bias permutation missingness approach to analyze non-monotone and non-ignorable missing data (Robins 1997; Robins et al. 1999). Unfortunately, the implementation of these approaches remains a challenging and computationally intensive task because of the lack of user-friendly routines.

4 Auxiliary variables

Effectiveness of the calibration and the propensity score approaches discussed so far depends crucially on the availability of a suitable set of auxiliary variables that help to explain both the nonresponse mechanism and the key study variables (see, e.g., Chapter 10 in Särndal and Lundström 2005). When survey data suffer from nonresponse errors, a distinction may be made between auxiliary variables that are observed only for the subsample of respondents (which we denoted by x) and auxiliary variables that are observed for all sample units (which we denoted by z). Such a distinction is important because the x-type variables can only be used as the benchmark variables in the calibration equations (as long as we also know the corresponding population total t_x), while the z-type variables can be used at any stage of the calibration and propensity score adjustments.

The most critical scenario occurs when the sampling frame contains very little auxiliary information for the implementation of nonresponse corrections so that we are forced to employ one-step calibration adjustments against U-level margins for basic x-type variables. Sometime useful z-

type variables can be obtained from paradata collected during the fieldwork such as interviewer characteristics, number of calls and indicators for nonresponse reasons. These variables have been frequently used as predictors of response propensity in the context of propensity score adjustments. For example, using paradata on the number of calls and indicators for whether the respondents ever refused, Kalton et al. (1990) and Meekins and Sangster (2004) show that respondents who are less cooperative in the first wave of a longitudinal study are more likely to become nonrespondents in the second wave. Similarly, Rizzo et al. (1994) and Loosveldt et al. (2002) find that amounts and patterns of item nonresponse in the first wave of a longitudinal study can be powerful predictors of attrition in the second wave.

For sample surveys using population registers as sampling frames, basic socio-demographic variables such as age, gender, and geographical indicators are typically observed for all sample units. This results in a more favorable scenario where the availability of some z-type variables makes it possible to perform nonresponse calibration adjustments against both S-level and U-level margins. The same variables may also serve as predictors of the unknown response propensities, but they do not have to coincide with the benchmark variables used in the calibration adjustments. As pointed out by Kott and Chang (2010), the distinction between these two sets of variables allows the treatment of nonignorable nonresponse as one can invoke a quasi-randomization approach to justify the nonresponse calibration adjustments.

In the most favorable scenario, there exists a large set of z-type variables. This is often the case in well designed sample surveys where the sampling frame and the data collection process provide valuable information on the key predictors of the response propensities and the key study variables of interest. Another instructive example is the adjustment for panel attrition in longitudinal surveys where one can typically exploit the additional information collected in previous waves. Notice that, in all these situations, the strategy of using as many as possible z-variables does not yield necessarily the 'best' nonresponse adjustment. First, models with a large set of auxiliary regressors may lead to greater bias than more parsimonious models (see, e.g., De Luca et al 2016 for a related discussion in the context of linear models). Second, with a view towards the MSE criterion, the inclusion of too many auxiliary regressors may also have a substantial impact on the precision of the estimated response propensities.

These considerations have given rise to a rapidly expanding literature on the issue of how to select a 'best' subset of auxiliary variables. For example, Rizzo et al. (1996) argue that the choice of auxiliary variables to account for panel attrition in longitudinal surveys could be more important

than the choice of the method to adjust the survey weights. The choice of a 'best' subset of auxiliary variables can be based on a variety of search algorithms, some of which are reviewed in Brick and Kalton (1996). Other relevant contributions in this direction include Schouten (2007) and Schouten et al. (2009), who suggest of choosing the best subset of auxiliary variables using similarity measures (the so-called R-indicators) between the respondents and the full sample. Särndal and Lundström (2010) proposed another set of indicators that account for the sample design, the set of observed respondents, and the specific calibration estimator. Comparisons with the approach by Schouten (2007) show that the two model selection algorithms do not always result in the same set of auxiliary variables.

The procedures used for preliminary screening of the available set of auxiliary variables may be subject to another issue, namely the fact that uncertainty introduced by the model selection step is not properly reflected into the final weighting system. In other words, if we use diagnostic tests to search for a best-performing model, then we need to take into account not only the uncertainty of the estimates in the selected model but also the fact that we have used the data to select a model. Model selection and estimation should be seen as a combined effort, not as two separate efforts, and failure to do so may lead to misleadingly precise estimates (see, e.g., Magnus and De Luca 2016). Problems associated with inference after model selection have been investigated in Magnus (1999, 2002), Leeb and Pötscher (2005, 2006, 2008), Berk et al. (2013), among others. All these studies suggest that ignoring the uncertainty associated with the model selection step can lead to seriously misleading inference. Another drawback of ignoring the noise produced by model selection is that small perturbations of the data may result in very different models being selected (Yang, 2001). From this perspective, the development of weighting methods that properly reflect the various forms of uncertainty arising in the construction and the adjustment of survey weights (e.g. calibration function, propensity score method, trimming, and auxiliary variables) remains a challenging, but extremely important, line of future research. A relevant contribution in this direction is given by Zigler and Dominici (2014), who account for model uncertainty in the context of propensity score estimation by using Bayesian variable selection and model averaging methods.

5 Conclusions

Nonresponse errors in sample surveys are nowadays the rule, not the exception, and one of the key issues is that the assumptions on the underlying nonresponse mechanism cannot be verified empirically. In this report we have provided a brief overview of the weighting methods available

for dealing with problems of unit nonresponse in cross-sectional surveys and problems of panel attrition in longitudinal surveys. The intuitive idea is to treat nonresponse as an additional - but unknown - phase of the sampling design and so that we then adjust the original sampling design weights of respondents to compensate for their systematic differences relative to nonrespondents. As usual, the validity of these weighting methods relies on the assumption that the nonresponse mechanism is MAR conditional on a set of benchmark variables.

We have focused our attention to the data producer perspective by discussing advantages and weaknesses of two broad classes of weighting methods: the calibration approach and the propensity score approach. Calibration adjustments are attractive because of their simplicity and because they allow us to align the weighted sample sums of some benchmark variables to the corresponding population values. The main drawback is that the choice of the calibration function corresponds to assuming an implicit parametric model for the nonresponse process and this feature of the calibration procedure cannot be robust to misspecification problems. Propensity score adjustments are in general more complex than calibration adjustments, but this extra effort is compensated by the possibility of relaxing strong parametric assumptions about the nonresponse process. To exploit the key advantages of both approaches, it is also possible to use a two-step procedure that involves some propensity score adjustment in the first step and some calibration adjustment in the second step.

Another key ingredient for successful weighting strategies is the auxiliary information available to compensate for nonresponse bias. Notice that the available auxiliary variables must be associated with both the response propensity and the key study variables. Using auxiliary variables that are weakly associated with the study variables of interest and the response propensity may only result in a lower precision of the weighted sample statistics. Also notice that when exploiting the auxiliary information as calibration margins we need to know the population totals and the auxiliary variables only for the subsample of respondents. In contrast, for the propensity score approach, we need to observe the auxiliary variables for both respondents and nonrespondents. These considerations make it clear that the collection of the good auxiliary information must be planned at the survey design stage as it involves choices regarding the sampling frame and the collection of paradata from the fieldwork.

After ensuring that enough auxiliary variables are available, the next challenges become how to select a best subset of auxiliary variables and how to reflect the uncertainty surrounding the construction of the survey weights. In recent years, similar problems have originated a rapidly

expanding literature on model selection and model averaging procedures, as well as a large literature on multiple imputation procedures for dealing with item nonresponse errors. To our knowledge, practical guidance on the development of weighting methods that account for problems of model uncertainty is still limited. More research in this direction is needed.

References

Brick, J. M. (2013). Unit nonresponse and weighting adjustments: a critical review. *Journal of Official Statistics* 29: 329–353.

Brick, J. M., and Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research* 5: 215–238.

Cohen, T., and Spencer, D. B. (1991). Shrinkage weights for unequal probability samples. *Proceedings of the section on survey research methods, American Statistical Association*, 625–630.

Da Silva, D. N. (2003). Adjustments for survey unit nonresponse under nonparametric response mechanisms. *Retrospective Theses and Dissertations - Iowa State University.* N. 1464.

Da Silva, D. N., and Opsomer, J. D. (2006). A kernel smoothing method of adjusting for unit non-response in sample surveys. *The Canadian Journal of Statistics* 34: 563–579.

Da Silva, D. N., and Opsomer, J. D. (2009). Nonparametric propensity weighting for survey nonresponse through local polynomial regression. *Survey Methodology* 35: 165–176.

De Luca, G., Magnus, J. R., and Peracchi, F. (2016). Balanced variable addition in linear models, with an application to the long-term health effects of childhood circumstances. *Mimeo.*

Deville, J. C., and Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* 87: 376–382.

Ekholm, A., and Laaksonen, S. (1991). Weighting via response modeling in the finnish household budget survey. *Journal of Official Statistics* 7: 325–337.

Eltinge, J. L., and Yansaneh, I. S. (1997). Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. Consumer Expenditure Survey. *Survey Methodology* 23: 33–40.

Han, P. (2014). Multiply robust estimation in regression analysis with missing data. *Journal of the American Statistical Association* 109: 1159–1173.

Hartley, H. O. (1946). Discussion of "A review of recent statistical developments in sampling and sample surveys". *Journal of the Royal Statistical Society* 109: 37–38.

Haziza, D., and Beaumont, J. F. (2007). On the construction of imputation classes in surveys. *International Statistical Review* 75: 25–43.

Haziza, D., and Lesage, E. (2016). A discussion of weighting procedures for unit nonresponse. *Journal of Official Statistics* 32: 129–145.

Hu, Z., Follmann, D. A., and Qin, J. (2012), Semiparametric double balancing score estimation for incomplete data with ignorable missingness, *Journal of the American Statistical Association* 107: 247–257.

Kalton, G., and Brick, M. (2000). Weighting in household panel surveys. In Advid, R. (Ed.) *Researching Social and Economic Change. The Uses of Household Panel Studies.* (p. 96–112). Routledge, New York.

Kalton, G., and Flores-Cervantes, I. (2003). Weighting Methods. *Journal of Official Statistics* 19: 81-97.

Kalton, G., Lepkowski, J., Montanari, G. E., and Maligalig, D. (1990). Characteristics of second wave non-respondents in a panel survey. *Proceedings of the section on survey research methods, American Statistical Association*, 462–467.

Kim, J. K., and Kim, J. J. (2007). Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics* 35: 501-514.

Kott, P. S., and Chang, T. (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association* 105: 1265–1275.

Lazzeroni, L. C., and Little, R. J. A. (1993). models for smoothing post-stratification weights. *Proceedings of the section on survey research methods, American Statistical Association*, 764–769.

Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory* 21: 21–59.

Leeb, H. and Pötscher, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *Annals of Statistics* 34: 2554–2591.

Leeb, H. and Pötscher, B. M. (2008). Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory* 24: 338–376.

Li, L., Shen, C., Li, X., et al. (2013). On weighting approaches for missing data. *Statistical Methods in Medical Research* 22: 14–30.

Little, R. J. A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review* 54: 139–157.

Loosveldt, G., Pickery, J., and Billiet, J. (2002). Item nonresponse as a predictor of unit nonresponse in a panel survey. *Journal of Official Statistics* 18: 545–557.

Lundström, S., and Särndal, C. E. (1999). Calibration as a standard method for treatment of nonresponse. *Journal of Official Statistics* 15: 305-327.

Magnus, J. R. (1999). The traditional pretest estimator. *Theory of Probability and Its Applications* 44: 293–308.

Magnus, J. R. (2002). Estimation of the mean of a univariate normal distribution with known variance. *Econometrics Journal* 5: 225–236.

Magnus, J. R., and De Luca, G. (2016). Weighted-average least squares (WALS): a survey. *Journal of Economic Surveys* 30: 117–148.

Meekins, B. J., and Sangster, R. L. (2004). Predicting wave nonresponse from prior wave data quality. *Proceedings of the section on survey research methods, American Statistical Association*, 4015–4021.

Oh, H. L., and Scheuren, F. J. (1983). Weighting adjustments for unit nonresponse. In Madow, W. G., Olkin, I., and Rubin, D.B. (eds). *Incomplete Data in Sample Surveys* V. 2 (p. 143–184). Academic Press, New York.

Phipps, P., and Toth, D. (2012). Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data. *Annals of Applied Statistics* 6: 772–794.

Politz, A., and Simmons, W. (1949). An attempt to get "not at homes" into the sample without callbacks. *Journal of the American Statistical Association* 44: 9–31.

Rizzo, L., Kalton, G., and Brick, M. (1994). Adjusting for panel nonresponse in the Survey of Income and Program Participation. *Proceedings of the section on survey research methods, American Statistical Association*, 422–427.

Rizzo, L., Kalton, G., and Brick, M. (1996). A comparison of some weighting adjustment methods for panel nonresponse. *Survey Methodology* 22: 43–53.

Robins, J.M., and Gill, R.D. (1997). Non-response models for the analysis of non-monotone ignorable missing data. *Stat Med* 16: 39–56.

Robins, J. M. (1997). Non-response models for the analysis of non-monotone non-ignorable missing data. *Stat Med* 16: 21–37.

Robins J. M., Rotnitzky, A., and Scharfstein, D. (1999). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In: Halloran, M., and Berry, D. (eds). *Statistical models in epidemiology: the environment and clinical trials* (p.1–92). Springer-Verlag, New York.

Rosenbaum, P. R., and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70: 41–55.

Rotnitzky, A., and Vansteelandt, S. (2015). Double-robust methods. In: Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., Verbeke, G. (eds). *Handbook of Missing Data Methodology* (p.185–209). CRC Press, New York.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika* 63: 581–590.

Schouten, B. (2007). A selection strategy for weighting variables under a not-missing-at-random assumption. *Journal of Official Statistics* 23: 51–68.

Schouten, B., Cobben, F., and Bethlehem, J. (2009). Measures for the representativeness of survey response. *Survey Methodology* 35: 101–113.

Särndal, C. E., and Lundström, S. (2005). *Estimation in surveys with nonresponse.* Wiley, Chichester.

Sändal, C. E., and Lundstroṁ, S. (2010). Design for estimation: identifying auxiliary vectors to reduce nonresponse bias. *Survey Methodology* 36: 131–144.

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

Skinner, C. J., and D'Arrigo, J. (2011). Inverse probability weighting for clustered nonresponse. *Biometrika* 98: 953–966.

Vermeulen, K., and Vansteelandt, S. (2015). Biased-reduced doubly robust estimation. *Journal of the American Statistical Association*, DOI:10.1080/01621459.2014.958155.

Yang, Y. (2001). Adaptive regression by mixing. *Journal of the American Statistical Association* 96: 574–588.

Zigler, C. M., and Dominici, F. (2014). Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects, *Journal of the American Statistical Association* 109: 95–107.