STATISTICS IN TRANSITION new series, March 2018
Vol. 19, No. 1, pp. 45–60, DOI 10.21307/stattrans-2018-003

JOINT RESPONSE PROPENSITY AND CALIBRATION METHOD

Seppo Laaksonen[1]**, Auli Hämäläinen**

ABSTRACT

This paper examines the chain of weights, beginning with the basic sampling weights for the respondents. These were then converted to reweights to reduce the bias due to missing quantities. If micro auxiliary variables are available for a gross sample, we suggest taking advantage first of the response propensity weights, and then of the calibrated weights with macro (aggregate) auxiliary variables. We also examined the calibration methodology that starts from the basic weights. Simulated data based on a real survey were used for comparison. The sampling design used was stratified simple random sampling, but the same methodology works for multi-stage sampling as well. Eight indicators were examined and estimated. We found differences in the performance of the reweighting methods. However, the main conclusion was that the response propensity weights are the best starting weights for calibration, since the auxiliary variables can be more completely exploited in this case. We also tested problems of calibration methods, since some weights may lead to unacceptable weights, such as below 1 or even negative.

Key words: reweighting, simulation study, macro vs. micro auxiliary variables, case of negative and other implausible weights

1. Introduction

Nonresponse and coverage problems are common in surveys. Both problems are increasing rather than declining. Unless the fieldwork is successful or special data collection modes are found and used, post-survey adjustments are the only option to try for improving the data quality. In this study, we concentrated on weighting adjustments. Reweighting is useless without appropriate auxiliary data. That is, we cannot do much without these variables. We tested two types of auxiliary variables: (i) aggregate, or macro, and (ii) micro.

Both types of auxiliary variables require that their values be available both for the respondents and for the non-respondents, and hopefully for ineligibles as well. In multi-stage designs the micro variables are often more difficult to get since the first stage is an area or an address, but macro variables still can be created. In the case of element-based sampling such as stratified simple random sampling, there are principle reasons favouring to try to get as good and many micro

[1] University of Helsinki Finland. E-mail: Seppo.Laaksonen@Helsinki.Fi.

auxiliary variables as possible to be used. On the other hand, there are many alternatives for macro variables. For instance, if a primary sampling unit (PSU) is an area cluster, it is possible to get various types of aggregate figures at this level. For instance, Laaksonen et al. (2015) could not get education at the micro level, but they had access to real grid data that gave the opportunity to construct the proportion of highly educated people at each grid. This is not, of course, as good as an education code with several categories at the micro level, but improved their weighting to some extent. In general, some types of calibration margins can always be used as macro auxiliary variables. This feature is a good reason to take advantage of calibration methods.

Brick (2013) presented an overview to weighting adjustments in the case of unit nonresponse. His earliest citations were from 1940's. He identified three major themes for nonresponse. Statistical adjustment of the survey weights to adjust for survey nonresponse is his third theme, while retaining the design-based mode of inference. He presented, at a general level, the following weighting methodologies: response propensity weighting, response homogeneity group weighting or weighting class methodology (see also Little and Rubin, 2002), propensity stratification (Valliant et al., 2013) and calibration estimation following Deville and Särndal (1992). Brick mentions also that post-stratification as a basic calibration estimator has been used for decades (Holt and Smith, 1979; Smith, 1991).

Post-stratification was augmented by Deville and Särndal (1992), leading to a more general approach so that several margins can be used to benchmark the reweights as precisely as the recent known population figures imply. This is the initial approach to calibration and can be used if the certain population margins are available. The quality of these margins should be as good as possible to succeed well in calibration. They are the types of benchmarking figures so that these 'estimates' will be automatically like 'true' values. This quality is not guaranteed for other estimates or for proper survey estimates, but it is expected that their bias will be reduced to some extent. The reason is that the benchmark margins correct often for frame and nonresponse errors. And more generally, it follows that an appropriate calibration method could possibly be an "ending method", after possible other weighting methods based on micro auxiliary variables. We follow this strategy.

Särndal and Deville worked later together with Sautory (1993) leading to a SAS macro Calmar. After that, the prosperity of the calibration methodology was ready to begin. Later, a second version of the SAS macro, Calmar 2, was published and became publicly available, providing new options for calibration (le Guennec and Sautory, 2005), including five distance functions. The macro gives opportunity to insert the margins of two levels (e.g. households and household members), but we do not examine this feature in this paper.

The distance function is used to minimize the change between the starting weights and the new calibrated weights while the benchmark margins are satisfied. The often applied distance function is linear, but this might be problematic since some weights can be negative or below one and this is not acceptable since the weight of every respondent should be at least one. Calmar 2 fortunately has four other functions, and two of them never yield to implausible weights; that is, raking ratio and sinus hyperbolicus, respectively. Two of these other methods include the bound option that may help in getting correct weights,

but it is not clear which bounds to use. It is also possible that the algorithm does not work with inappropriate bounds. In general, the use of bounds is usually subjective, and the target is to get acceptable weights, but it is not guaranteed that they are reducing the bias in estimates. Valliant et al. (2013) say that these can be moved to the boundary, but we do not agree with this statement since it is even more subjective. We thus do not recommend subjective strategies in weighting or other survey methodologies.

Calibration methodology flourished extensively in the 2000's, and various specifications were developed (Kott, 2006; Kott & Chang, 2008, 2010; Lumley et al., 2011; Särndal, 2007). However, linear calibration was the common method. The often used form of it was the generalized linear regression estimation method GREG (Estevao and Särndal, 2006; Särndal, 2007; Henry and Valliant, 2015; Valliant et al., 2013).

Another approach to reweighting is to exploit micro level auxiliary variables as well as possible. The basic ideas of this methodology are mainly from the 1980's (Little 1986). Little and Rubin (2002) use the term "propensity weighting" that we also use, but adding the word "response", which was used by Brick (2013). The model behind the response propensity (RP) weighting is usually logistic regression, but probit regression (Laaksonen and Heiskanen, 2014) and other link functions can be applied as well. First applications of RP weighting were done at the group level, often called response homogeneity groups, adjustment cells or weighting classes (Valliant et al., 2013; Brick, 2013; Little and Rubin, 2002; Ekholm and Laaksonen, 1991). The group methods are still much used (Haziza and Lesage, 2016).

Laaksonen (2007) applied the RP weighting technique so that he first estimates a logistic regression model for predicting the response propensities to the individual respondents. In the second stage, he divides the basic sampling weights with these propensities to get the preliminary weights. Finally, he benchmarks these weights to correspond to the known population of the explicit strata. This was done since the sampling design was the stratified random sampling and these population figures were available in the beginning, before the fieldwork. The success of this methodology depends on the richness of the micro-level auxiliary variables. Macro variables can be used, and they are usually predicting the missing quantities as well. The benchmarking in this study was ensured at the stratum level, or at the post-stratum level, if applied after post-stratification (Laaksonen et al., 2015).

This study combined both approaches, that is, calibration methods and response propensity weighting, which were not mentioned in Brick (2013) or in the textbook of Valliant et al. (2013). Our approach consisted of the three steps. First, the basic sampling weights were computed using the sampling design of the survey and assuming that the response mechanism was ignorable within strata, but not between strata. Then the response propensity weights were constructed, and finally, these weights were used as the starting weights in the calibration. The strategy gave more benchmarks than the initial RP weighting does. It took advantage of both micro and macro auxiliary variables that were available. We compared the different methods with each other using a simulated data set that was based on a modification of a real data set, but is less complex than the initial data set (Laaksonen and Heiskanen, 2014).

This artificial population was extracted from the data set of the about 3000 respondents and then copied enough times to get the universe of about 180 thousand. The missing indicators remained in the data, but some randomness was added in survey variables to get a more realistic file. This new universe gave the opportunity to see how well each method works in nearly real-life situations. We also compared variations of calibration methods that were invented in Calmar 2.

In Section 2, we present both the calibration methods of Calmar 2, our response propensity weighting method and the principles of the joint response propensity and the calibration method. Section 3 goes on to describe our simulated data. The following section summaries our empirical results, which clearly show that our main method was best on average, and never worst, even though it did not lead to essential improvement in all cases. Section 5 continues analysing the calibration weights of Calmar 2 with a partially new sample, which illustrates variation between distance functions. This analysis also exposed that both linear and logistic distance functions may bear unacceptable weights, and such weights cannot be used in practice without manipulation. The final section provides our conclusions.

The bias of mean estimates only is considered in the empirical part. We follow Brick and Jones (2008) in this sense. They said that variability can be measured reasonably well, whereas bias is difficult to measure due to nonresponse.

2. Calibration, response propensity weighting and their joint method

Successful calibration methods were developed in the early 1990's, when the Deville and Särndal article (1992) was published. The methods were further developed when the first Calmar SAS macro was coded by the French statistical office INSEE in 1993 (Deville et al., 1993; Sautory, 2003; Willenberg, 2009). The new version, Calmar 2, published in 2003, offered new resources for performing calibrations and implements the generalized calibration method of handling non-response.

The theory of calibration estimators takes advantage of a distance function between the starting weight and the new calibrated weight, $G(w_k/d_k) = G(x_k) = G(x)$. This distance should be minimized while the desired calibration margins are satisfied. These margins are vectors of the macro auxiliary variables given by the user. The calibrated weights yield these same "estimates", thus concerning aggregates of auxiliary variables. It does not ensure anything about the accuracy of survey estimates. Some improvement is expected if the margins correct for the frame errors, for example. The results are expected to be less biased if the macro auxiliary variables and the survey variables are correlated.

Calmar 2 is a SAS macro into which a user can choose the three types of options:

(i) the initial or starting weight, d_k, which will be calibrated (we have two alternatives for this weight),
(ii) the calibration margins that are expected to be as true population totals as possible,
(iii) the calibration methods with alternative distance functions.

Calmar uses Lagrange multipliers in minimizing the distance function $x=d_k/w_k$, in which w_k is the calibrated weight. The original version of Calmar offered four calibration methods and later one more was offered (Le Guennec and Sautory, 2005; Mc Cormack, 2006), corresponding to different distance functions. This number is more than in most other software packages that usually mention the two first methods (Brick and Jones, 2008; Valliant et al., 2013). On the other hand, many other distance functions are mentioned in theoretical papers (Plikusas and Pumputis, 2010). The Calmar 2 methods are characterized by the form of function as follows:

- the *linear* method (the formula is $G(x) = \frac{1}{2}(x-1)^2$): the calibrated estimator is the generalized regression estimator,
- the *exponential* method (the formula is $G(x) = x\log \mathrm{x} - \mathrm{x} + 1$): this is also called the *raking ratio method*,
- the *logistic* method (the formula is $G(x) = xlog\frac{x}{1-x}$): this method provides lower limits *L* and upper limits *U* on the weight ratios *x* (bounds),
- the *truncated linear* method, in which the distance function is linear, but the bounds are included as in logistic method, and
- the sinus hyperbolicus method, which (the formula $G(x) = \frac{1}{2\alpha}\int_1^x \sinh\left[\alpha\left(t-\frac{1}{t}\right)\right]dt, \quad \alpha > 0 \quad implemented\ in\ CALMAR\ 2)$ does not give negative or other implausible weights.

Implausible weights are not ensured in the cases of the linear and the logistic methods without any bounds. However, it should be noted that a user should make the selection of these limits, maybe subjectively. We do not recommend using the bounds for this reason.

There are other tools to calibrate weights, but Calmar 2 is, to our knowledge, one of the best ones and used extensively in Europe. It includes several methods as well. Lumley's (2013) *R* package has been used much as well. For instance, the first nonresponse weights of the European Social Survey (2014) were produced using this *R* package. These are called post-stratified weights even though they are like the raking-ratio weights of Calmar 2.

For this paper, we obtained results with all five methods, but we present the detailed simulation results only with the linear method. This is due to fairly similar results obtained with all five methods, and hence these minor differences are not interesting. The reason for the minor differences was our ordinary sampling design (Section 4). On the other hand, we continued with a specific sampling design in Section 5, which illustrates better problems obtained with linear and logistic calibration. Moreover, this design and its sample also illustrate the role of bounds. We used similar bounds in all empirical analysis so that they are symmetric, relatively. This means that the upper bound was equal to 5 and the lower bound was its inverse, that is 0.2=1/5**.** The range of these was ordinary. Using limits, the algorithm failed to converge in one case (see Section 5).

The strategy for creating 'response propensity weights' was as follows (Laaksonen and Heiskanen, 2014):

(i) We obtained the gross sample design weights that are the inverses of the inclusion probabilities. These inclusion probabilities varied by strata but were constant within each stratum.

(ii) We assumed that the response mechanism within each stratum is ignorable (assuming that the response mechanism is random within strata but not random between strata), and hence computed the basic weights analogously to the weights (i). These are available only for the respondents k, and symbolised by $w_k = \frac{N_h}{r_h}$, in which N_h refers to the target population, r to the respondents and h to four strata.

(iii) Next, we took those basic weights and divided them by the estimated response probabilities (called also response propensities) of each respondent obtained from the logit (probit link gives quite similar results) model, and symbolised by p_k. It is good to concentrate on the model building if several auxiliary variables are available, e.g. interactions can be tried.

(iv) Before going forward, it is good to check that the probabilities p_k are realistic, that is they are not too small (let say below 0.05), for instance. All probabilities were, of course, below 1, and hence all weights were plausible.

(v) Since the sum of the weights (iii) did not match the known population statistics by strata h, they should be calibrated so that the sums are equal to the sums of the basic weights in each stratum. This was done by multiplying the weights (iii) by the ratio $q_h = \frac{\sum_h w_k}{\sum_h w_k / p_k}$. This is one option for the response propensity modelling weighting, called "Pure" in Table 2.

(vi) It is good also to check these weights against basic statistics, such as the mean, the maximum, the minimum and the coefficient of variation. This was done for the first sample and as soon as the weights gave plausible results, the next repetitions were performed in the same way.

The joint response propensity and calibration (JRPC) weighting means that we used the response propensity weights as the starting weights in calibration, whereas these are the basic weights in pure calibration. This study is focused on the joint method, but we compared all other weights in the same framework. JRPC weighting is a two-stage calibration method. It is possible to perform it in one stage as well. Kott & Chang (2008, 2010) present such a method, but it is not as straightforward as our solution to work with both methods. The methodology by Haziza & Lesage (2016) follows the similar strategy that combined response propensity weighting and linear calibration. Their simulation application is 'fictional' (not from a real case) and hence it is difficult to see how well comparable it is with our framework. On the other hand, it is not possible to see details of their response propensity model that are of high importance in practice.

We combined both methods, but each stage can be considered separately. We think that the two-stage strategy is rational. It is not even necessary to continue to calibrate if 'a client' is happy with the RP weighting. Much depends on the availability of good calibration margins. It is, however, good to remember that the calibration as an ending method is useful if the calibration margins are correct. These margins are often known well by users (e.g. distribution by gender or age group) and if they are not correct, survey estimates may not be trusted, either, even though these are not necessarily related to each other.

3. Data and simulation principles

The data for simulations were created from the 2010 Finnish Security Survey (Laaksonen and Heiskanen, 2014) so that its three independent data sets from the respondents (face-to-face, phone and web) were first pooled together. Then this data set was extended from about 3,000 respondents to the artificial target population data set with 180,000 people. The extension was rather straightforward, but minor randomness was added to income values, among others, to avoid same values. As far as the absence of the target population was concerned, we followed as well as possible the initial unit nonresponse, and hence the response rate of our data was about equal and about 49%.

We expended much effort in our initial study to gather as many auxiliary variables as possible. So, we had the same chance to use these in our simulations as well. Since the simulation sample was essentially smaller than in the initial survey, we had to apply a bit less demanding model. However, our final response propensity model consisted of the following explanatory variables (number of categories in parenthesis): interaction of gender (2) and age group (5), education level (6), stratum (4), partnership (2), children or not at home (2), unemployed or not (2), mother tongue (3), number of rooms of house (4), if living in the municipality born or not (2). These were not very significant in all samples, but a good point in response propensity-based adjustments is that insignificancy does not violate the adjusted weights, but its impact on an estimate is less remarkable in such a case. Thus, it may not improve the estimates in all simulation samples.

The absence itself was very randomized, but it had a similar feature as in the initial survey. There was thus an absence indicator for each target population unit (Brick and Jones, 2008). When drawing samples from this population, absence varied in each sample correspondingly. This added uncertainty to the simulations. The response of each sample followed the Bernoulli scheme, that is, the number and distribution of the respondents (and by strata) varied randomly to some extent.

There are the types of variables that do not exactly correspond to those of the appendix of the paper by Laaksonen and Heiskanen (2014). The violence variables were based on 8 to 10 binary questions as to whether a respondent has met that violence problem at least once. Then, the prevalence indicator was estimated. The income variable was continuous, and it was expected to be explained better than the other seven using the auxiliary variables available that were the same as in the published paper.

Table 1 shows the averages of all indicators in the entire population. In simulations, we try to get the estimates that are as close to these values as possible. It is not easy for many reasons. One special reason is that there can be a rather small number of observations for some indicators. This is indicated in the right column. This absence is not due to nonresponse, it is mainly due to the topic itself. For example, if a person never had a partner, the answer is empty. The reason for some absence is not exactly known, such as violence by stranger recently or harassment ever. On the other hand, we do not need to know absences well, we only need to calculate the estimates and compare these to those true target population figures. When interpreting the results, it is however good to keep in mind that some estimates are computed from a small sample size.

Table 1. Major statistics for 15-79 years old population. 179,985 persons in the simulations. The response rate below 99% means that the question did not concern them

Indicator in simulations	Average in population	Response rate, %
Income (yearly)	44905€	100
Worry (about crime)	28%	100
Harassment recently	43%	99
Harassment ever	74%	24
Violence by stranger recently	33%	24
Violence by stranger ever	87%	41
Violence by partner	16%	74
Violence by ex-partner	30%	45

The simulation strategy, naturally, followed the survey principles:

(i) Four explicit strata by four large regions were formed.
(ii) A simple random sample with a disproportional allocation was drawn from each stratum, and altogether equalled 2,000 individuals. The disproportionality was moderate (the maximum 3 times as big as the minimum). This simple design meant that paying attention to side effects due to complex sample design was not needed.
(iii) Basic sample weights were computed for the respondents as usual, dividing the target population sizes by the number of the respondents (assuming ignorable unit non-response).
(iv) The calibrated weights were computed using Calmar 2 from the basic weights. The margin variables were four strata, two genders and five age groups. These variables are quite easily available in many countries. In principle, we could add more margins, but this was not realistic since it is possible in a few cases in practice only, and would require more resources. We had more margins in our specific study (Section 5).
(v) Response propensity (RP) weights were respectively calculated.
(vi) The similar calibration as in (iv) was performed taking the RP weights as the starting weights.
(vii) The mean estimates were calculated using all weights.
(viii) The procedure from (ii) to (vii) was repeated 150 times and the output data set obtained.
(ix) The results between simulated results and true values were compared.

4. Summary of the simulation results

We drew 150 samples from this simulation population using stratified simple random sampling, which is the most common design in countries with a population register frame. This number of simulations is not big. One reason is that the whole Calmar procedure was fairly demanding, and took time while the outputs were not easy to handle further. The second reason is that the estimates were found to be stable, even after 70 simulations. Two indicators, however, did not become very stable. These were “violence due to stranger ever” and ”violence due to partner”. These results should be interpreted with caution. If the difference is minor between the two methods compared, it should not be taken seriously. This point is not critical to our simulation study, but it is general in surveys with enough complex estimates. The estimates thus are not always ideal even using good weights.

The relative bias from the true value (=(estimate-true value))/true value) is the most illustrative way to compare results since it is not needed to look at the indicator values themselves (their averages are in Table 1). It is common in other studies such as Brick and Jones (2008). The comparisons are presented in Table 2.

Table 2. Results for the relative bias from the true value (%) with basic weights and RP weights, and both continued with linear calibration. The term “Pure” means without calibration. The most biased estimates are in red, the best ones are bolded. The order of the indicators is by the success of the joint RP and linear calibration (last column). The standard error of simulations is small (from 0.1% to 0.5%)

Indicator	Basic weight		Response propensity	
	Pure	Calibration	Pure	Calibration
Violence by ex-partner	**0.60**	-2.72	-0.83	-1.21
Harassment ever	-1.36	-2.22	-0.53	**-0.42**
Worry	0.76	-0.84	0.16	**-0.02**
Violence by stranger recently	-1.03	**-0.10**	0.12	0.15
Harassment recently	6.91	0.55	0.62	**0.32**
Income	2.06	1.79	0.39	**0.33**
Violence by partner	7.24	**4.22**	4.68	4.52
Violence by stranger ever	6.50	**2.39**	4.86	5.27
Average success ranking by four methods (1=best, 2 =second best, 3=third best, 4=worst)	3.38	2.50	2,25	**1.88**

The last row of Table 2 shows an overall ranking of the methods. If the ranking was interpreted straightforwardly, we see that the best method was joint response propensity and calibration, with pure response propensity being the second and pure basic weighting the worst. There are exceptions, nevertheless. Pure basic weighting was best for “violence by ex-partner”, for which any method

does not work well. Pure calibration worked well for another difficult indicator, "violence by partner". This indicator seemed to be most difficult to estimate well with any method. We cannot explain why the joint method was as bad with this method, but better than obtained by the basic weights.

Another hard indicator to estimate correctly was "violence by stranger ever". The auxiliary variables were not appropriate to predict absence if any method was not reasonably good. Fortunately, we see that this succeeds well with some indicators, the best being worry, then income, violence by stranger recently and then harassment recently. In these cases, basic weights did not lead to reliable estimates. These weights were created without other auxiliary variables except region that is used in sampling design. It was well understood that the bias in income can be reduced using micro auxiliary variables available in our study. It was interesting that the similar reduction was found in worry as well. Our joint method even gave slightly better results than pure RP weighting.

In general, almost all weights with adjustments improved the estimates to some extent, but they were either upward or downward biased. Secondly, it seems that even sophisticated weights did not always improve the accuracy substantially. A good point is that they did not deteriorate them, either, although the improvement was minor. It is good to keep in mind that the calibration as the ending method was good if the margins were "true population totals". Respective estimates, such as income aggregates, were obviously more reliable as well.

5. Testing possibility to get inappropriate weights

We tested five weights, although our simulation results in Table 2 concerned only linear calibration. The estimates by the different calibration weights are approximately equal. This is due to our sampling design, in which the sample allocation into strata was fairly proportional (Section 3) and the response mechanism was same as earlier. Our auxiliary variables in calibration were also of good quality. For these reasons, all of our weights were correct, at least in the sense that their values were above one.

However, it was realized that some weights may lead to implausible weights that are below one or even below zero. We did not find references with empirical examples that examined this problem, although it is well known (the problem was mentioned in Deville and Särndal, 1992). Hence, we wanted to test this awkward opportunity with our simulation data. The linear weights are most well-known problematic weights. The implausible weights may be avoided using the bounds given for the Calmar 2 macro. These bounds are the relative limits of the calibrated weights compared to the starting weight, thus a lower, and the upper bound, respectively. The bounds naturally are given subjectively. After some attempts, we decided to choose the following limits: LOW=0.2, UP=5. The range is rather ordinary, but we failed in all experiments to get any result. These limits are possible to give for the two distance functions, both for linear and logistic, but we do not recommend using such bounds. Raking ratio and sinus hyperbolicus never give implausible weights.

Our special experiment included the modification for our basic simulations. One is the sample size that was reduced with 50 percent (from 2000 to 1000) but the same absence indicator was used as in simulations. The sample allocation

was made more disproportional. Interestingly, the minimum gross sample design weights were close to each other, and decreased from 36 to 31. It is good to recognize that a small weight can more easily remove below 1 unless very strict limits in bounds are used. On the other hand, a strict limit may mean that the algorithm fails to converge.

The second difference is that we created more calibration margins. Moreover, we tested two types of auxiliary margins, those derived from the target population frame and those derived from one sample. The latter is often used in practice, since true margins are not always possible to get. It has been supposed that it does not matter much if the sample data do not fit well to the real-life population. For example, the European Social Survey (2014) used margins for post-stratified weights that were derived from the Eurostat labour force survey although its quality is not complete. This data source was the best available and, hence, it was used.

We tested the three types of auxiliary margins presented in Table 3, all based on both the real population and the sample data.

We knew in advance that it was possible to get implausible weights either with linear calibration or logistic calibration, and hence we did not take care of other calibration methods, although they were used. Table 4 gives the summary of linear calibration. The weights were calculated similarly as in simulations.

Table 3. The margins used in the specific examination, obtained from the simulation data (true values) or from one sample data

(i)	The same as in our simulations, thus gender, age group and region but using 13 age groups (instead of 5 in the simulation).
(ii)	Adding first education with five categories
(iii)	Adding then marital status with four categories.

As Table 3 shows, all margins required more than in the first simulations since the number of age groups was essentially larger. This did not damage the weights when starting from basic weights, or even when the margins were not ideal, that is, when using sample-based margins. The negative weights were not met while the three margins were correct and the weights were response propensity-based. Instead, starting from the response propensity weights and from sample margins, negative weights were received. This problem worsened when the number of calibration margin variables was increased.

Table 4. Amount of invalid weights in linear calibration. Margins, see Table 3.

Starting weights	Negative weights, per cent	
	Population margins	Sample margins
Basic weights, margins (i)	0.0	0.0
RP Adjusted weights, margins (i)	0.0	7.8
Basic weights, margins (ii)	6.2	6.6
RP Adjusted weights, margins (ii)	3.7	9.7
Basic weights, margins (iii)	7.9	13.0
RP Adjusted weights, margins (iii)	7.8	11.0

Table 5 summaries the results of incorrect weights using logistic calibration. This method only can give weights below 1, but not negative weights. These problems were less dramatic than in the case of linear weighting, but however they may occur. The special case is that the Calmar 2 algorithm could not get any weights if the margins were sample-based and the calibration was started from RP weights. We found that this is more common if the bounds are stricter than we used.

Table 5. Amount of invalid weights in logistic calibration with bounds. N.A. means that the calibration algorithm did not converge

Starting weights	Weights below 1, per cent	
	Population margins	Sample margins
Basic weights, margins (i)	0.0	0.0
RP Adjusted weights, margins (i)	0.0	0.0
Basic weights, margins (ii)	0.0	0.0
RP Adjusted weights, margins (ii)	1.4	1.6
Basic weights, margins (iii)	0.0	0.0
RP Adjusted weights, margins (iii)	3.7	N.A.

What to do if the weights are below one? Our recommendation is not to increase these weights subjectively above one, but to change the calibration strategy. There are several options to do so. The best one is to use another distance function, but it is possible also to collapse calibration margins. Naturally, it is good to use as good margins as possible, since sample-based margins might be inconsistent. This topic should be further examined.

6. Conclusion

This study compared four weighting methods so that one group of these methods was either the basic weighting or response propensity weighting. The second group was calibrated so that either the basic weights or the response propensity weights were used as the starting weights for calibration. The calibration of the first applications was following the linear distance function that works technically in the first group when the number of the calibration margin variables is three and they are true values. In this case when the number of calibration categories is moderate, and they are true values, the impact of the distance function was not big, that is, the estimates were about equal. This pattern does always give plausible weights, but that is not the case if more calibration margin categories are added.

If the calibration margins are not true population values, there is a danger that implausible weights can be obtained. This was tested in this paper in the second part, using distance functions other than linear. When the calibration margins were drawn from the sample, some implausible weights were found. Our data set was not simple, as is often the case in real-life. Our purpose was not to get the ideal estimates only, but those that are realistic. Our eight indicators were used in exercises and hence well illustrated the situation in survey practice. Two of these

eight indicators were found to be difficult to estimate well due to the lack of good macro and micro auxiliary variables. Fortunately, we found that the reweights for the rest of other indicators substantially reduced the bias. It should be noted that the data environment was demanding and indicators concerning crimes due to violence even more so. Their prevalence was hard to examine in surveys, in general, due to their sensitivity, and it was expected that the weights would not help much. The drawback is the reality that the outcome depends on the respondents, since the weights can only use such data; if certain groups are represented to a small extent among the respondents, the weights cannot help.

We examined easier variables, such as income and worry due to crime, and the reweights helped much more. We cannot make any straightforward conclusion about the weights applied, however our study shows that the combination of the response propensity weighting and calibration is the best of all four methods. This takes advantage both of micro and macro auxiliary variables. The first ones were especially used in response propensity weighting and the second ones in the calibration performed from these first weights. This two-stage strategy is not often used, but we recommend it. It is definitely better than the often used pure calibration with linear distance function. The calibration is good to be used as the ending method since it ensures that estimates derived from the known population margins are correct.

We conducted tests with linear calibration and found that it works correctly if the variation of the starting weights is moderate, the number of margins is not big and the margins are accurate. In other cases, linear calibration may lead to negative weights. It is also possible that logistic calibration may give the weights below one; if the bounds used are strict, the algorithm of the method does not always converge. Further investigations with weighting adjustments are needed. Special attention could be paid to get good predicting auxiliary variables with a high quality, both at micro and macro level.

REFERENCES

BRICK, J. M., (2013). Unit Nonresponse and Weighting Adjustments: A Critical Review, Journal of Official Statistics, 29, 3, pp. 329–353.

BRICK, J. M., JONES, M. E., (2008). Propensity to respond and Nonresponse Bias, METRON – International Journal of Statistics, LXVI, 1, pp. 51–73.

DEVILLE, J-C., SÄRNDAL, C-E., (1992). Calibration Estimators in Survey Sampling, Journal of the American Statistical Association, pp. 376–382.

DEVILLE, J-C., SÄRNDAL, C-E., SAUTORY, O., (1993). Generalized Raking Procedures in Survey Sampling, Journal of the American Statistical Association, pp. 1013–1020.

EKHOLM, A., LAAKSONEN, S., (1991). Weighting via Response Modelling in the Finnish Household Budget Survey, Journal of Official Statistics. 7.2, pp. 325–337.

ESS _ European Social Survey, (2014). Documentation of ESS Post-Stratification Weights.

http://www.europeansocialsurvey.org/docs/methodology/ESS_post_stratification_weights_documentation.pdf.

ESTEVAO, V. M., SÄRNDAL, C-E., (2006). Survey Estimates by Calibration on Complex Auxiliary Information, International Statistical Review 74, 2, pp. 127–147.

HAZIZA, D., LESAGE, E. (2016). A Discussion of Weighting Procedures for Unit Nonresponse, Journal of Official Statistics 32, 1, pp. 129–145, http://dx.doi.org/10.1515/JOS-2016-0006.

HENRY, K. A., VALLIANT, R., (2015). A design effect measure for calibration weighting in single stage samples, Survey Methodology 41, 2, pp. 315–331.

HOLT, D., SMITH, T. M. F., (1979). Post-Stratification. Journal of the Royal Statistical Society, Series A (General). Vol. 142, pp. 33–46.

KOTT, P., (2006). Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors. Survey Methodology 32, 2, pp. 133–142.

KOTT, P. CHANG, T., (2008). Can calibration be used for 'nonignorable' nonresponse, Proceedings of Joint Statistical Meeting. Section of Survey Research, pp. 251–260.

KOTT, P., CHANG, T., (2010). Using Calibration Weighting to Adjust for Nonignorable Unit Nonresponse, Journal of the American Statistical Association, pp. 105–491.

LUMLEY, T., (2013). Survey: Analysis of Complex Survey Samples. R package version 3.29, URL http://CRAN.R-project.org/package=survey.

LUMLEY, T., SHAW, P., DAI, J. Y., (2011). Connections between Survey Calibration Estimators and Semiparametric Models for Incomplete Data. International Statistical Review 79, 2, pp. 200–220, DOI: 10.1111/j.1751-5823.2011.00138.x.

LAAKSONEN, S., (2007). Weighting for Two-Phase Surveyed Data. Survey Methodology, December Vol. 33, No. 2, pp. 121–130, Statistics Canada.

LAAKSONEN, S., (2015). Sampling Design Data File, Survey Statistician 72, pp 61–66.

LAAKSONEN, S., HEISKANEN, M., (2014). Comparison of Three modes for a Crime Victimization Survey, Journal of Survey Statistics and Methodology, 2 (4), pp. 459–483, DOI 10.1093/jssam/smu018.

LAAKSONEN, S., KEMPPAINEN, T., STJERNBERG, M., KORTTEINEN, M., VAATTOVAARA, M., LÖNNQVIST, H., (2015). Tackling City-Regional Dynamics in a Survey Using Grid Sampling. Survey. Research Methods by the European Survey Research Association, Vol 9, No 1, pp. 55–65, www.surveyresearch.org.

Le GUENNEC. J., SAUTORY. O., (2005). CALMAR 2: Une Nouvelle Version de la Macro Calmar de Redressment D'Échantillion Par Calage, http://vserver-insee.nexen.net/jms2005/site/files/documents/2005/327_1-JMS2002_SESSION1_LE-GUENNEC-SAUTORY_CALMAR-2_ACTES.PDF.

LITTLE. R. J. A., (1986). Survey Nonresponse Adjustments for Estimates of Means, International Statistical Review, 54, pp. 139–157.

LITTLE, R. J. A., RUBIN, D. B., (2002). Statistical Analysis with Missing Data, 2nd Edition. Wiley.

LUNDQUIST, P., SÄRNDAL, C-E., (2013). Aspects of Responsive Design with Applications to the Swedish Living Conditions Survey. Journal of Official Statistics 29, 4, pp. 557–582, DOI: org/10.2478/jos-2013-0040.

LUNDSTRÖM, S., SÄRNDAL, C-E., (1999). Calibration as a Standard Method for Treatment of Nonresponse, Journal of Official Statistics, 15, 2, pp. 305–327.

McCORMACK, K., (2006). The calibration software CALMAR – What is it? Central Statistics Office Ireland, http://vesselinov.com/CalmarEngDoc.pdf.

ROBINS, J. M., ROTNITZKY, A., ZHAO, L.-P., (1994). Estimation of regression coefficients when some regressors are not always observed, Journal of the American Statistical Association 89, pp. 846–866.

PLIKUSAS, A., PUMPUTIS, D., (2010). Estimation of the finite population covariance using calibration. Nonlinear Analysis: Modelling and Control, Vol. 15, 3, pp. 325–340.

SAUTORY, O., (2003). CALMAR 2: A New Version of the Calmar Calibration Adjustment Program, Proceedings of Statistics Canada's Symposium: Challenges in Survey Taking for the Next Decade.

SÄRNDAL, C-E., (2007). Calibration Approach in Survey Theory and Practice, Survey Methodology, 33, No. 2, pp. 99–119.

SMITH, T. M. F., (1991). Post-stratification, The Statistician 40, pp. 315–323.

VALLIANT, R., DEVER, J. A., KREUTER, F., (2013). Practical Tools for Designing and Weighting Survey Samples, Statistics for Social and Behavioral Sciences. Springer.

WITTENBERG, M., (2009). Sample Survey Calibration: An Information-theoretic perspective, A Southern Africa Labour and Development Research Unit Working Paper Number 41, Cape Town: SALDRU, University of Cape Town.