TEST

Comments on: Deville and Särndal's Calibration: revisiting a 25 years old successful optimization problem

--Manuscript Draft--

Manuscript Number:	
Full Title:	Comments on: Deville and Särndal's Calibration: revisiting a 25 years old successful optimization problem
Article Type:	Discussion
Corresponding Author:	María del Mar Rueda Universidad de Granada SPAIN
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Universidad de Granada
Corresponding Author's Secondary Institution:	
First Author:	María del Mar Rueda
First Author Secondary Information:	
Order of Authors:	María del Mar Rueda
Order of Authors Secondary Information:	
Funding Information:	

Noname manuscript No.
(will be inserted by the editor)

Comments on: Deville and Särndal's Calibration: revisiting a 25 years old successful optimization problem

Maria del Mar Rueda

the date of receipt and acceptance should be inserted later

1 Introduction

It is a pleasure to comment on this interesting article by Devaud and Tillé, whose paper gives us the opportunity to reflect on the important impact produced by this method of sampling parameter estimation. Since the seminal work by Deville and Särndal, calibration has been one of the most useful tools available with which to incorporate auxiliary information in survey sampling. This technique ensures that the estimates obtained are coherent with those already in the public domain, while simultaneously reducing non-coverage, non-response and selection biases ([1]). Although other important estimation methods that also use auxiliary information have been proposed (e.g. the empirical likelihood method; [7] or that of model-based estimators; [29]), in practice, the vast majority of national statistical agencies use calibration as a reweighting technique and have developed software to compute calibrated weights in accordance with the auxiliary information available in administrative records and other reliable sources.

The authors have reformulated the classical calibration problem, focusing on issues related to optimisation. In this commentary, I shall focus on certain topics that were not discussed in the paper and that involve modifications to the optimisation problem.

2 Model-calibrated estimation

The use of calibrated estimators obtained from the constraint $\sum_{k\in s} \omega_k \mathbf{x}_k = T_{\mathbf{x}}$ is motivated by a linear model. If another type of curve is employed to relate the study variable y with the auxiliary variables $\mathbf{x}$, the calibration estimators

M. Rueda
Department of Statistics and O.R. University of Granada
E-mail: mrueda@ugr.es

considered to date may be ineffective.

The model-calibrated estimators suggested by [30] are based on the superpopulation model: $E_\xi(y_k|\mathbf{x}_k) = \mu(\mathbf{x}_k, \theta)$ where θ is an unknown population parameter, $\mu(\mathbf{x}, \theta)$ is a known function of $\mathbf{x}$ and θ, and E_ξ denotes the expectation with respect to the superpopulation model. Under this model-assisted approach, the auxiliary information should be used in accordance with the values $\mu(\mathbf{x}_k, \widehat{\theta})$ $k = 1, \ldots, N$ where $\widehat{\theta}$ is an estimator of θ. The model-calibration estimator of T_y is defined as $\widehat{T}_{YMC} = \sum_{k\in s} \omega_k y_k$ but the weights ω_k, minimise a distance function subject to the new constraints:

$$\sum_{k\in s} \omega_k = 1 \qquad \sum_{k\in s} \omega_k \mu(\mathbf{x}_k, \widehat{\theta}) = \sum_{k\in U} \mu(\mathbf{x}_k, \widehat{\theta}).$$

[30] showed that under certain conditions the estimator is asymptotically unbiased under the design (irrespective of whether the model is correct or not) and also that it is approximately unbiased under the model. These authors also obtained an expression for the asymptotic design variance. $\widehat{T}_{YMC}$ matches the conventional calibration estimator when a linear model is used.

The model-assisted approach has facilitated the use of new regression techniques in the calibration environment. Thus, [21] formulated calibration estimators considering regression models with a random component. In recent years, studies have also considered the use of ML algorithms for deriving model-assisted estimators (see e.g. [4]). However, little research has been undertaken in which these ML methods are applied to obtain calibrated estimators. The papers by [19] and [20], which use neural networks and p-splines to formulate non-parametric model-calibrated estimators, are important exceptions in this area. In a recent simulation study, [9] compared model-calibrated estimators based on various machine learning methods.

3 Calibration for other parameters

In section 2, the authors discuss calibration with respect to the estimation of the population total. The calibration technique can be used to estimate more complex parameters than a population total but the constraint must be modified to match the specific parameter to be estimated. Some examples are presented below.

[12] considered a general case, where the parameter of interest θ_N is the solution to an estimation equation of the form $H(\theta) = \sum_{k\in U} h(x_k, y_k, \theta) = 0$ for $h(\cdot)$ a continuous differentiable function of θ. The calibration estimator of θ_N is a solution to equation $\hat{H}(\theta) = \sum_{k\in s} w_k h(x_k, y_k, \theta) = 0$ where w_k are determined from $\sum_{k\in s} \omega_k \mathbf{x}_k = T_{\mathbf{x}}$. These authors obtained the asymptotic properties of the estimator.

A case of special importance is the estimation of the finite population distribution function (fdf) and its application to the estimation of quantiles and poverty measures. The non-smooth character of the fdf creates certain complexities that have been resolved in different ways. Thus, [11] proposed an estimator of the distribution function based on a calibration constraint specified by the quantile of an auxiliary variable. [23] considered a pseudo-variable $g_k = \widehat{\beta}' \mathbf{x}_k$ for $\widehat{\beta}$, the weighted estimator of the regression coefficient, and obtained new calibrated weights subject to the calibration equations $\frac{1}{N}\sum_{k\in s}\omega_k \Delta(t_j - g_k) = \frac{1}{N}\sum_{k\in U}\Delta(t_j - g_k) \quad j = 1, 2, \ldots, P$. Under some conditions, the estimator $\widehat{F}_{yc}(t)$ is a genuine distribution function. [15],[16] and [17] addressed the problem of selecting the optimal auxiliary vector included in the constraints. In a similiar way, [18] used a global penalised calibration method to define a new estimator for the fdf. Other related papers include [30], [24] and [2].

Interesting papers on the question of calibration estimation for other non-linear estimators have been presented by [26] and [22].

4 Functional calibration

An interesting extension of calibration emerges when both the survey and the auxiliary variables are considered as random functions. In this respect, [6] proposed a Horvitz Thompson estimator of the mean trajectory, and under certain assumptions regarding the sampling design, a functional central limit theorem has been proposed. Based on this estimator, [8] extended the approach taken to obtaining calibration sampling weights using functional data.

Assuming there exists a unique functional random variable $Y_i(t)$ that is observed for the unit in the sample s and also that the population mean of an auxiliary functional vector is available, the functional estimator for the population mean $\bar{Y}(t)$ based on functional calibration weights is expressed by the linear weighted estimator $\frac{1}{N}\sum_{k\in s} w_k(t)Y_k(t)$, where the new weights satisfy the set of constraints $\frac{1}{N}\sum_{k\in s}\omega_k(t)\mathbf{X}_k(t) = \bar{\mathbf{X}}(t)$. This functional calibration constraint is expressed as an infinite-dimensional linear inverse problem, the solution to which is derived by means of the maximum entropy on the mean principle.

5 Calibration in online surveys

In recent decades, new technologies have had a profound impact on survey techniques. This impact has been observed in social and political surveys, and particularly in market research surveys, where the greater speed and reduced costs obtained via new technologies have changed the ways in which data are collected. The use of web surveys is currently expanding at a rapid pace. Nevertheless, estimates obtained from this source present three significant issues: the absence of a sample design enabling the use of conventional methods of inference in finite populations, the existence of frame coverage problems and the danger of non-response bias. Although many statistical methods to reduce non-coverage and non-response bias have been proposed, the problems arising from the non-randomness of the sample are more complex and have yet to be satisfactorily addressed. As a result, the use of web surveys can result in unreliable inferences being drawn.

Several methods can be used to correct selection bias in online surveys (see e.g. [5] or [28]). Calibration can also be applied to remove the bias in non-probabilistic samples but its application is not immediate. There are no sample weights and therefore it is necessary to define which weights d_k must be considered in the distance function to be minimised. One straightforward solution is to take $d_k = N/n$, that is, to consider the sample of volunteers as if it were obtained with a simple random sampling design. When analysing online survey results, it is reasonable to assume that the decision to take part depends on the respondents characteristics and that the probability of their participating varies from one person to another. In this respect, [3] showed that bias can be reduced through calibration only when non-response due to volunteering has a missing at random scheme, and that this approach is not valid in other situations (which are the most frequent).

In the last two decades, propensity score adjustment (PSA) has been increasingly used as a promising means of correcting selection bias in online surveys. In this approach, a reference probabilistic survey is conducted on the same target population as that of the online survey, and used to estimate the probability of participation. Research has shown that PSA successfully removes bias in some situations, but at the cost of increasing the variance ([14]). [27] showed that the estimation of a variable using PSA must be complemented with further weighting adjustments in order to reduce the bias in the resulting estimates. The use of PSA with further calibration has been studied by [14] and [10], who concluded that calibration adjustments are helpful if they are applied using appropriate covariates.

Acknowledgements This study was partially supported by Ministerio de Economía y Competitividad (Grant MTM2015-63609-R).

References

1. Arcos, A., Contreras, J.M., Rueda, M. (2014) A novel calibration estimator in social surveys. Sociological Methods and Research 43(3), 465–489
2. Arcos, A., Martínez, S., Rueda, M., Martínez, H. (2017) Distribution function estimates from dual frame context, Journal of Computational and Applied Mathematics 318, 242–252.
3. Bethlehem, J.: Selection Bias in Web Surveys. Int. Stat. Rev., 78(2), 161-188 (2010)
4. Breidt, J., Opsomer, J.: Model-assisted survey estimation with modern prediction. Stat. Sci., 32 (2), 190-205 (2017)
5. Buelens, B., Burger, J., van den Brakel, JA. (2018) Comparing Inference Methods for Non-probability Samples. Int Stat Rev 86(2):322343.
6. Cardot, H., Josserand. J. (2011). Horvitz-Thompson estimators for functional data: asymptotic confidence bands and optimal allocation for stratified sampling. Biometrika, 98, 107118.
7. Chen J. and Qin J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. Biometrika 80, 107–116.

8. Gallón, S. Gamboa, F, Loubes, J.M. (2015) Functional calibration estimation by the maximum entropy on the mean principle, Statistics, 49(5), 989–1004.
9. Ferri, R., Castro, L., Rueda, M. (2019). Superpopulation models using Machine Learning methods for estimation in online surveys. In Proceeding of Mathematical and Computational Modelling, Approximation and Simulation. MACMAS 2019.
10. Ferri, R., Rueda, M. (2018). Efficiency of Propensity Score Adjustment and calibration on the estimation from non-probabilistic online surveys SORT 42(2),159–182
11. Harms, T., Duchesne, P. (2006). On calibration estimation for quantiles, Survey Methodology 32, 37–52.
12. Kim, J.K. and Park, M. (2010). Calibration estimation in survey sampling. International Statistical Review 78, 21–39.
13. Kovacevic, M. (1997) Calibration estimation of cumulative distribution and quantile functions from survey data. Proceedings of the Survey Methods Section, Statistical Society of Canada 47, 139–144.
14. Lee, S., Valliant, R. (2009). Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment. Sociol. Method. Res. 37(3), 319-343.
15. Martínez, S., Rueda, M., Arcos, A., Martínez, H. (2010). Optimum calibration points estimating distribution functions. Journal of Computational and Applied Mathematics 233(9, 2265–2277.
16. Martínez, S., Rueda, M., Martínez, H., Arcos, A. (2015) Determining P optimum calibration points to construct calibration estimators of the distribution function. Journal of Computational and Applied Mathematics 275, 281–293.
17. Martínez, S., Rueda, M., Martínez, H., Arcos, A.(2017). Optimal dimension and optimal auxiliary vector to construct calibration estimators of the distribution function. Journal of Computational and Applied Mathematics 318, 444–459.
18. Mayor Gallego, J.A., Moreno Rebollo, J.L., Jiménez Gamero, M.D. (2019). Estimation of the finite population distribution function using a global penalized calibration method. AStA. Advances in Statistical Analysis 103, 1–35.
19. Montanari, G.E., Ranalli, M.G. (2005). Nonparametric model calibration estimation in survey sampling. Journal of the American Statistical Association, 100, 1429–1442.
20. Montanari, G. E., Ranalli, M. G.(2009). Multiple and ridge model calibration. In Proceedings of Workshop on Calibration and Estimation in Surveys. Statistics Canada.
21. Park, M. and Fuller, W.A. (2009). The Mixed Model for Survey Regression Estimation.Journal of Statistical Planning and Inference 139, 1320–1331.
22. Plikusas, A. (2006). Non-linear calibration, in Proceedings, Workshop on survey sampling. Venspils, Latvia. Riga: Central Statistical Bureau of Latvia.
23. Rueda, M. and Martínez, S. and Martínez, H. and Arcos, A.(2007). Estimation of the distribution function with calibration methods. J. Statist. Plann. Inference, 137(2),435–448.
24. Rueda, M., Sánchez-Borrego, I., Arcos, A., Martínez, S. (2010). Model-calibration estimation of the distribution function using nonparametric regression. Metrika 71, 33-44.
25. Särndal C.E. (2007). The calibration approach in survey theory and practice. Survey Metholodogy 33(2), 99-119.
26. Singh, S. (2001). Generalized calibration approach for estimating variance in survey sampling. Annals of the Institute of Statistical Mathematics 53, 404–417.
27. Valliant, R., Dever, J. A. (2011). Estimating Propensity Adjustments for Volunteer Web Surveys. Sociol. Method. Res. 40(1), 105–137.
28. Valliant, R. Comparing alternatives for estimation form nonprobability samples. Technical report. 2019. DOI: 10.13140/RG.2.2.14618.90566
29. Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). Finite Population Sampling and Inference. A Prediction Approach. John Wiley. New York.
30. Wu, C., and Sitter, R.R.(2001) A model-calibration approach to using complete auxiliary information from survey data. Journal of the American Statistical Association 96 (453), 185–193.