Towards a Global Convergent Algorithm for Integer Calibration Weighting

Kelly Toppin* Luca Sartore† Clifford Spiegelman‡

Abstract

The USDA's National Agricultural Statistics Service (NASS) conducts the U.S. Census of Agriculture in years ending in 2 and 7. The census describes the characteristics of U.S. farms and the people who operate them. To adjust for under-coverage, nonresponse and misclassification, NASS produces the weights on the responding records using a capture-recapture methodology. However, the weights need to be further refined through a calibration process so that the census estimates agree with known population values. The current algorithm (called INCA) was developed to provide integer calibrated weights per NASS requirements. In INCA, weights adjusted for undercoverage, nonresponse, and misclassification are first rounded using an optimal rounding procedure, and then integer programming using coordinate descent is performed on the integer weights. However, the existence of multiple local solutions makes the search of a global solution exponentially complex. This article describes a comparison between two algorithms for integer calibration based on an L1-norm relative error. The results of a simulation study designed to investigate the properties of the estimator is presented.

Key Words: DSE, Weights, Global integer optimization, Census of Agriculture

1. Introduction

USDA's National Agricultural Statistics Service (NASS) conducts the U.S. Census of Agriculture every five years, in years ending in 2 and 7, to provide useful information about the U.S. farms, ranches, and the people who operate them. The Census is also the only source of uniform comprehensive agricultural information for every state and county in the United States.

The Census is based on the NASS list frame that consists of agricultural operations, some of which do not satisfy the farm definition (see O'Donoghue et al., 2009, for further details). The update of the list frame is an ongoing process. The Census Mailing List (CML) is a "frozen" image of the list frame at a specific point in time. The CML is incomplete. To address the resulting undercoverage, NASS turned to a Dual-System Estimation (DSE) methodology to adjust the estimates not only for under-coverage, but also for non-response and incorrect classification of farms. The adjustments are expressed as record weights (DSE weights). However, these are computed without taking into account the information from reliable administrative sources. Thus, the application of a calibration weighting technique is necessary to assure that consistent estimates are produced across all levels of aggregation.

Weighting calibration methods can provide a better set of weights for the Census, for which known totals from administrative data and other trusted sources are available. These calibrated weights are associated with their corresponding records to account for under-coverage, non-response, misclassification, and other fluctuations from known totals. A two-step approach is commonly adopted to calculate the necessary adjustments. First, the Census weights are initially adjusted to compensate for under-coverage, non-response

*National Agricultural Statistics Service, United States Department of Agriculture, 1400 Independence Ave. SW, Washington, DC 20250, kelly.toppin@nass.usda.gov

†National Institute of Statistical Sciences, 19 T.W. Alexander Drive, P.O. Box 14006, Research Triangle Park, NC 27709-4006

‡Texas A&M University, 3135 TAMU, College Station, TX 77843-3135

and/or misclassification. Second, calibration is applied to further improve the weights and produce unbiased estimates.

The first concept of weighting calibration was introduced by Lemel (1976). Deville and Särndal (1992) further developed the idea by providing methods that modify the design weights in the Horvitz-Thompson estimator (Horvitz and Thompson, 1952). Singh and Mohl (1996) heuristically justified the use of calibration and provided some algorithms that were compared to each other. Théberge (1999) provided some calibration techniques based on linear algebra properties to estimate some parameters of the population other than totals and means. Duchesne (1999) developed a robust method that satisfies the calibration benchmarks while forcing range restrictions on the weights. Théberge (2000) also studied the impact of weight restrictions and discussed the effects between outliers and extreme weights. Estevao and Särndal (2000) proposed the "functional form method" to the calibration problem to remove the limitation due to the minimization of a distance measure between the weights by satisfying the calibration benchmarks expressed in a "functional form". Calibration successively expanded into the realm of adjustments for nonresponse and coverage errors (Kott, 2006).

Integer calibrated weights are based on a standard that NASS adopted for its U.S. Census of Agriculture publication reports. This concept allows NASS to produce more consistent tabulations where the estimates sum to the correct totals across all levels of aggregation. Scholetzky (2000) investigated the effects of rounding the weights instead of total estimates. Integer calibration (INCA) was developed at NASS for producing the final weights of the U.S. Census of Agriculture (see Sartore et al., 2018, for further details on the algorithm). This methodology has a limitation: INCA converges to a local minimum that may not be a global solution.

From the authors' experience, there are two alternative ways to produce integer calibrated weights:

1. rounding the DSE weights that are successively calibrated by dealing only with integer weights, or

2. performing calibration with constrained real weights that will be rounded according to optimality criteria.

In this article, the second approach is considered and developed to produce an efficient algorithm to be compared with INCA.

The next section describes an attempt of a globally convergent algorithm that produces integer calibrated weights. A case study is presented in Section 3, while concluding remarks are summarized in Section 4.

2. Methodology

While DSE weights account for nonresponse, misclassification, and undercoverage, they do not automatically provide integer totals. Calibration weighting methods change the values of the adjusted weights produced by a DSE approach. These values are usually collected on a vector $\mathbf{w}^*$, and are calibrated to match known totals. Producing integer calibrated weights $\hat{\mathbf{w}}$ that are "close" to the DSE weights $\mathbf{w}^*$ is equivalent to solving the following optimization problem:

$$\min_{\mathbf{w}\in\mathcal{N}\subset\mathbb{N}^n} \delta(\mathbf{w}-\mathbf{w}^*) + \lambda\rho(\mathbf{y}-\mathbf{A}\mathbf{w}), \tag{1}$$

where

n denotes the number of observations or units responding to the Census,

$\mathbf{y}$ is a vector of m known totals,

$\mathbf{A}$ is an $m \times n$ matrix of collected data,

$\delta(\cdot)$ represents a function based on a distance measure defined as $\delta : \mathbb{R}^n \to \mathbb{R}_+$,

$\rho(\cdot)$ denotes a loss function based on a distance measure defined as $\rho : \mathbb{R}^m \to \mathbb{R}_+$,

λ is a positive scalar that controls the importance of the errors produced by the benchmark equations $\mathbf{y} = \mathbf{A}\mathbf{w}$.

2.1 Calibration

To solve the integer calibration problem in (1), a similar problem based on real numbers can be set as

$$\min_{\mathbf{w} \in \mathcal{W} \subset \mathbb{R}^n} \delta(\mathbf{w} - \mathbf{w}^*) + \lambda \rho(\mathbf{y} - \mathbf{A}\mathbf{w}),$$

where the bounded set $\mathcal{W}$ is defined as

$$\mathcal{W} = \bigotimes_{i=1}^{n} [1, u_i],$$

where $\otimes$ denotes the Cartesian product among the sets $[1, u_i]$, with $u_i > 1$ is the upper bound of the i-th weight, for any $i = 1, \ldots, n$.

For a generic approach, gradient descent algorithms can establish a base for an algorithm that converges linearly to a global solution by forcing the search on $\mathcal{W}$. The transformation

$$w_i = \frac{1 + u_i \exp(x_i)}{u_i + u_i \exp(x_i)} \tag{2}$$

is considered to simplify the search; therefore, an optimal solution can be found by performing the optimization in an unbounded setting:

$$\min_{\mathbf{x} \in \mathbb{R}^n} \delta(\omega(\mathbf{x}) - \mathbf{w}^*) + \lambda \rho(\mathbf{y} - \mathbf{A}\omega(\mathbf{x})),$$

where the function $\omega : \mathbb{R}^n \to \mathbb{R}^n$ performs the transformation (2).

The improved Rprop$^+$ algorithm (Igel and Hüsken, 2003) is used to decrease iteratively the objective formulated in (1). The algorithm updates the weights according to criteria that are based on the sign of the derivatives computed in two successive steps.

The algorithm starts by computing the gradient as

$$\mathbf{g}^{(\tau)} = \mathbf{J}(\mathbf{x}) \left[\nabla \delta\{\omega(\mathbf{x}) - \mathbf{w}^*\} - \lambda \mathbf{A}^\top \nabla \rho\{\mathbf{y} - \mathbf{A}\omega(\mathbf{x})\} \right],$$

where the diagonal matrix $\mathbf{J}(\mathbf{x}) = \text{diag}(\theta_1, \ldots, \theta_n)$, with

$$\theta_i = \frac{u_i \exp(x_i)(u_i - 1)}{\{u_i + u_i \exp(x_i)\}^2}, \ \forall i = 1, \ldots, n.$$

Two constants are set such that $0 < \eta^- < 1 < \eta^+$, and a vector $\mathbf{\Delta}^{(0)}$ is initialized by setting its components as

$$\Delta_i^{(0)} = \begin{cases} \Delta_-, & \text{if } d_i < \Delta_-, \\ \Delta_+, & \text{if } d_i > \Delta_+, \\ d_i, & \text{otherwise,} \end{cases}$$

where $\Delta_- < \Delta_+$ are both non-negative constants, $d_i = |g_i^{(\tau)}| / \max(|g_{(1)}^{(\tau)}|, |g_{(n)}^{(\tau)}|)$, with $g_{(1)}^{(\tau)}$ and $g_{(n)}^{(\tau)}$ are respectively denoting the minimum and the maximum value of the gradient at the step τ.

At the step $\tau = 0$, the weights are updated as

$$x_i^{(\tau+1)} = x_i^{(\tau)} - \text{sign}(g_i^{(\tau)})\Delta_i^{(\tau)}, \tag{3}$$

but for $\tau > 0$, the algorithm computes the gradient $\mathbf{g}^{(\tau)}$ and it updates the vector $\mathbf{\Delta}^{(\tau)}$ according to the changes of the corresponding partial derivatives. At this point, the algorithm considers three cases:

- when $g_i^{(\tau)} g_i^{(\tau-1)} < 0$. The values of $\mathbf{x}$ are adjusted only if the objective at the step $\tau - 1$ is lower than the objective at the current step, so that

  $$x_i^{(\tau+1)} = x_i^{(\tau)} - \text{sign}(g_i^{(\tau)})\Delta_i^{(\tau-1)}.$$

  The size of the step is then updated as $\Delta_i^{(\tau)} = \max(\Delta_i^{(\tau-1)}\eta^-, \Delta_-)$, and $g_i^{(\tau)} = 0$;

- when $g_i^{(\tau)} g_i^{(\tau-1)} > 0$. The step size is adjusted as $\Delta_i^{(\tau)} = \min(\Delta_i^{(\tau-1)}\eta^+, \Delta_+)$, and the values of $\mathbf{x}$ are updated as in (3);

- when $g_i^{(\tau)} g_i^{(\tau-1)} = 0$. No adjustment of the size step is require and the values of $\mathbf{x}$ are updated as in (3).

The procedure terminates its iterations when any among the following convergence criteria is satisfied:

- the decrement of the objective function is insignificant;
- $\Delta_i = \Delta_-$, for all $i = 1, \ldots, n$;
- τ is greater than a fixed number of iterations.

A better solution can be achieved by starting the optimization from distinct initial candidate solutions and taking the best point as the final solution. The rounding algorithm is performed on the vector of calibrated weights $\tilde{\mathbf{w}}$ to obtain integer values.

2.2 Rounding algorithm

The rounding algorithm applies rounding rules as a stochastic search inspired by evolutionary algorithms that heuristically converge to the best rounded values that provide an optimal solution to the problem (4).

The algorithm starts by taking the lower integer of each calibrated weights so that $\check{w}_i = \lfloor \tilde{w}_i \rfloor$. These values are used to compute optimal binary steps $s_i \in \{0, 1\}$ that will be used to compute the optimal solution as $\hat{w}_i = \check{w}_i + s_i$ by optimizing the following problem:

$$\min_{\mathbf{s} \in \{0,1\}^n} \delta(\check{\mathbf{w}} - \mathbf{w}^* + \mathbf{s}) + \lambda\rho(\mathbf{y} - \mathbf{A}\check{\mathbf{w}} - \mathbf{A}\mathbf{s}). \tag{4}$$

The gradient of this objective is computed as

$$\mathbf{g} = \nabla\delta(\check{\mathbf{w}} - \mathbf{w}^* + \mathbf{s}) - \lambda\mathbf{A}^\top\nabla\rho(\mathbf{y} - \mathbf{A}\check{\mathbf{w}} - \mathbf{A}\mathbf{s}).$$

The adaptive strategy updates the vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, which are initialized as

$$\begin{aligned}\boldsymbol{\alpha}^{(0)} &= \mathbf{1} - \tilde{\mathbf{w}} - \check{\mathbf{w}} - \mathbf{v} + 2\mathbf{s}^*,\\ \boldsymbol{\beta}^{(0)} &= \mathbf{1} - \tilde{\mathbf{w}} + \check{\mathbf{w}} + \mathbf{v} + 2(\mathbf{1} - \mathbf{s}^*),\end{aligned}$$

where $\mathbf{1}$ is a vector of ones, the components in the vector $\mathbf{v}$ are defined as

$$v_i = \begin{cases} 0, & \text{if } g_i \geq 0 \text{ or } \check{w}_i + 1 > u_i,\\ g_i/g_{(1)}, & \text{otherwise,}\end{cases}$$

and $\mathbf{s}^*$ is a suboptimal solution for the problem in (4) obtained with deterministic methods (e.g. see Sartore et al., 2018). Whenever $\mathbf{s}^*$ is not available, let it be a vector where its components are set to 0.5.

At each iteration the components of possible candidate solutions $\tilde{\mathbf{s}}_k$ are simulated from a $\mathsf{Bernoulli}(p_i)$, where $p_i = \alpha_i/(\alpha_i + \beta_i)$, for $k = 1, \ldots, K$, where K denotes the total number of simulations to perform. For $\tau > 0$, the values of the two vectors are updated as

$$\begin{aligned}\boldsymbol{\alpha} &\leftarrow \boldsymbol{\alpha} + \sum_{k\in\mathcal{K}} \tilde{\mathbf{s}}_k,\\ \boldsymbol{\beta} &\leftarrow \boldsymbol{\beta} + \sum_{k\in\mathcal{K}} (\mathbf{1} - \tilde{\mathbf{s}}_k),\end{aligned}$$

where $\mathcal{K}$ is the set of the indexes, which have binary steps $\tilde{\mathbf{s}}_k$ that achieve the most reduction of the objective.

This algorithm converges when p_i tends to zero or one for all $i = 1, \ldots, n$, or τ exceeds a fixed number of iterations to perform.

The optimal values of the steps are obtained as

$$\hat{s}_i = \begin{cases} 0, & \text{if } \alpha_i < \beta_i,\\ 1, & \text{otherwise.}\end{cases}$$

3. Case study

The performance of the proposed algorithm is studied through simulations, and the results are compared to those produced by the integer calibration algorithm developed by Sartore and Toppin (2016). The values of 150 weights ω_i are drawn from a $\mathsf{Gamma}(3.333, 1)$ distribution, and the data stored in the components of a 201×150 matrix $\mathbf{A}$ satisfy the following equality:

$$a_{ji} = \begin{cases} 1, & \text{if } j = 1,\\ b_{ji}c_{ji}, & \text{otherwise,}\end{cases}$$

where b_{ji} is drawn from a $\mathsf{Bernoulli}(0.3)$ distribution and c_{ji} from a $\mathsf{Poisson}(4)$, for any $j = 2, \ldots, 201$ and $i = 1, \ldots, 150$. The calibration benchmarks are successively computed as a system of linear equations, i.e. $y = A\omega$, while the DSE weights are drawn from a $\mathsf{U}(0, 7.5)$ distribution to increase the difficulty of attaining an optimal solution to the calibration problem. This means that more operations are performed to obtain a solution that satisfies NASS requirements, i.e. $\hat{w}_i \in [1, 6]$, for any $i = 1, \ldots, 150$.

Table 1: Results of the simulation study

	$\lambda = 0.09$	$\lambda = 0.37$	$\lambda = 1.49$	$\lambda = 5.97$
Objective	910.43	3102.69	11894.86	47097.84
MAD	1.16	1.10	1.07	1.07
TAE	7901.00	7873.00	7862.00	7862.00
Time (s)	44.52	44.08	47.41	44.01

For this study, a simplified objective function was designed to reduce the distance of the population totals from the calibration benchmarks, i.e.

$$\rho(y - Aw) = \sum_{j=1}^{201} \left| y_j - \sum_{i=1}^{150} a_{ji} w_i \right|,$$

which has a straightforward evaluation of its gradient, i.e.

$$g_i = -\sum_{j=1}^{201} \text{sign}(\varepsilon_j) a_{ji},$$

where the error $\varepsilon_j = y_j - \sum_{i=1}^{150} a_{ji} w_i$, and the function

$$\text{sign}(z) = \begin{cases} -1, & \text{if } z < 0, \\ 0, & \text{if } z = 0, \\ 1, & \text{if } z > 0. \end{cases}$$

To evaluate the performance of the algorithm, it was applied using four values of λ, in particular $\lambda \in \{0.09, 0.37, 1.49, 5.97\}$. The optimizations all started from the same initial DSE weight vector. The final value of the objective function, the Total Absolute Error (TAE) from the calibration benchmarks, the Mean Absolute Deviation (MAD) from the DSE weights, and the computational cost expressed in elapsed seconds were all evaluated (see Table 1).

INCA produces a vector of integer weights with MAD at 1.84 and total absolute error of 1992.

4. Summary

From the data collection through the publication of the US Census of Agriculture, specific deadlines are set for each step of the process. Thus, it is not possible to adopt and execute computational intensive algorithms within the time available. However, improvements can be made within a limited amount of additional time for a particular task. Thus, the goal is to provide a set of methods to compute the most precise estimates within the available time.

The proposed calibration algorithm adjusts the weights on an unbounded space and successively produces integer weights with the desired characteristics. However, the computational efficiency is dramatically reduced by performing several optimizations starting from different initial weights each time. This approach is necessary to allow for a more accurate search of a global solution. Under specific conditions on the objective function (Koenker, 2005) the optimal solution is unique and the developed algorithm heuristically converges to a vector of integer calibrated weights that coincides with the global solution.

Even if the methodology based on resilient back-propagation overcomes the main limitation of a discrete coordinate descent algorithm, INCA still finds better vectors of integer

calibrated weights within a small amount of time. This also provides more evidence to the fact that calibration performed by dealing with integer weights surpasses the achievements of the most efficient optimization algorithms for real-valued weights.

Future research can speed-up the minimization process and exploit a combined swarm-optimization when the objective function presents many local minima.

Acknowledgments

The authors would like to thank Nathan Cruze, and Linda Young for reviewing the earlier versions of this paper and providing useful comments. This research was supported by the intramural research program of the U.S. Department of Agriculture, NASS. The findings and conclusions in this preliminary publication have not been formally disseminated by the U.S. Department of Agriculture and should not be construed to represent any agency determination or policy.

References

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382.

Duchesne, P. (1999). Robust calibration estimators. *Survey Methodology*, pages 43–56.

Estevao, V. M. and Särndal, C.-E. (2000). A functional form approach to calibration. *Journal of Official Statistics*, pages 379–399.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, pages 663–685.

Igel, C. and Hüsken, M. (2003). Empirical evaluation of the improved rprop learning algorithms. *Neurocomputing*, 50:105–123.

Koenker, R. (2005). *Quantile regression*. Cambridge University Press, New York.

Kott, P. (2006). Using calibration weights to adjust for nonresponse and coverage errors. *Survey Methodology*, 32:133–142.

Lemel, Y. (1976). Une gnralisation de la mthode du quotient pour le redressement des enqułtes par sondage. *Annales de l'ins*, (22/23):273–282.

O'Donoghue, E., Hoppe, R. A., Banker, D., and Korb, P. (2009). Exploring alternative farm definitions: implications for agricultural statistics and program eligibility. *Economic Information Bulletin-USDA Economic Research Service*, 49.

Sartore, L. and Toppin, K. (2016). *inca: Integer Calibration*. R package version 0.0.2.

Sartore, L., Toppin, K., Young, L., and Spiegelman, C. (2018). Developing integer calibration weights for Census of Agriculture. *Journal of Agricultural, Biological and Environmental Statistics*, Accepted.

Scholetzky, W. (2000). Evaluation of integer weighting for the 1997 Census of Agriculture. Technical Report RD-00-01, United States Department of Agriculture, National Agricultural Statistics Service, Washington, DC.

Singh, A. and Mohl, C. (1996). Understanding calibration estimators in survey sampling. *Survey Methodology*, 22(2):107–115.

Théberge, A. (1999). Extension of calibration estimators in survey sampling. *Journal of the American Statistical Association*, pages 635–644.

Théberge, A. (2000). Calibration and restircted weights. *Survey Methodology*, pages 99–107.